FINAL REPORT

Author: W.F. Mackey

Title: Mechanolinguistic method analysis.

Div: VIII-C Report no. 2

MECHANOLINGUISTIC METHOD ANALYSIS

Report presented to the Royal Commission on Bilingualism and Biculturalism

W.F. Mackey
with the assistance of M.S. Mepham

June 1966

PREFACE

The purpose of this project was to devise a quick and effective technique for analysing language teaching methods and materials in use in Canada.

Since the analysis was to be as objective and as complete as possible all measurable variables had first to be quantified. This quantification necessitated the analysis of a volume of date too great to be processed by hand; it was therefore evident that the work would have to be done with aid of a computer.

Because the same process of analysis would have to be used for each method, the same series of operations being repeated again and again, computer analysis of methods -- mechanolinguistic method analysis -- was obviously the most appropriate and economical way of obtaining a quick, objective and complete analysis of such a large amount of language teaching material.

Fortunately, much work toward a quantitative analysis of methods had already been done; it remained to translate this into computer programs and to produce a prototype of mechanolinguistic analysis. This was the expressed objective of the research project. It was felt, however, that an isolated prototype, no matter how perfect, would not give a sufficiently clear idea of the sort of comparisons made possible by the analysis. It was therefore decided, at the risk of delaying the final report, to put a second method through the computer analysis

already developed for the first and to present both results side by side.

For the first method analysed by computer, we selected the French course most widely used in the Canadian Civil Service, viz., Voix et Images de France (Method a); for the second, we chose a comparable audio-visual course which had been used in certain government departments, viz., French through Pictures (Method b).

Before any of this material could be put into the computer, three types of computer programs had first to be elaborated--data control programs, language analysis programs and method analysis programs.

The data control programs were based on available language statistics for spoken French including those of the Centre de recherches pour la diffusion du français, which supplied some yet-unpublished frequency figures. This material was included in the data control programs, the purpose of which was to supply a category and numerical value for each item in the language likely to be found in the method.

The language analysis programs were based on the most formal procedures of analysis available for French. They included an automatic grammar which had to be specially devised for the purpose.

The method analysis programs were elaborated from parameters established in an earlier study (1961) published in London in 1965 (Language Teaching Analysis. London: Longmans).

Completion of these three types of mechanolinguistic programs was complex and time-consuming. Much time had to be spent in the preparation of such things as frequency dictionaries, structural grammars, and data searching sequences, in the perfection of component programs by various stages of approximation, the redesigning of procedures within the limited memory capacity of the computer, etc. When this complex of programs had been completed, each of its stages had to be verified, and ofter re-written, until all stages of analysis were sufficiently perfect and well-co-ordinated to produce an accurate prototype analysis of a single method. When this was achieved, it was a relatively simple matter to produce the analysis of a second method, since we had now reached the stage where out automatic method analysis was functioning.

As a result, it is now possible to make a rapid, automatic and detailed objective analysis of all methods and materials used in Canada for the teaching of the second language.

The complex of mechanolinguistic method analysis programs now available is the result of carefully co-ordinated team-work. The computuer programs were the work of Michael Mepham, as were the graphic representations of the results. Much of the work on the data control material was done by Lorne Laforge, Jean-Guy Savard, Jean-Marie Courtois, Flore Gervais, Monique Benoit, Pierre Cardinal, Gerald McNulty and Michèle Crevière. The technical assistance of Louis Robichaud, Pierre Ardouin and Roger Miville-Deschênes also contributed to the success of this venture.

W.F. M.

Table of Contents

List of Figures

List of Tables

1. Introduction

The choice of language teaching methods has always been a matter of opinion rather than of fact. Departments of education have up to now simply relied on the opinions of language teachers. These persons, often excellent classroom teachers with good judgment, could nevertheless not be aware of all the possible methods nor did they have the techniques of analysis to enable them to pass an objective judgment on those which came to their attention.

Every year thousands of learners are consequently introduced to the study of the second language through methods which are not the most suitable. In some provinces, the methods have first been "experimented" in classes where students are reported to have done "someone's idea of well on someone's idea of a valid test."

The failure of experimental learning situations in language teaching method analysis is now recognized as being due to the multiplicity of factors in the method, the teachers and the learners. Only after each complex of factors has been isolated, analysed and quantified can a technique be involved for determining with any certainty the most suitable language learning method for any given group.

The most important and the most neglected in the analysis of these three components is the method itself. Only after all the relevant factors in a method have been analysed and quantified, is it possible to determine the extent to which each relates to any teaching or learning situation.

The techniques for isolating and quantifying these factors in the analysis of methods had already been elaborated before the project began. They were grouped according to four fundamental pedagogical questions:

1. What elements are taught? (selection).
2. When are they introduced? (gradation).
3. How are they introduced? (presentation).
4. How are they exercised? (repetition).

A number of the measurements devised could be automated; that is, to say, effected on a digital computer. We proposed to reduce human intervention in the execution of the analysis while producing a more detailed and rigourous description than could be undertaken manually.

2.- The Description of the Method

Of the possible variables, only those which belong to one or more of the pedagogical factors are considered. Of these, some are more easily evaluated manually. For the purposes of the description, the variables which can be measured automatically are grouped under the factor they most effectively describe.

2.1.- Selection

The value of the method depends on the number of different elements, the nature of these elements and their utility within the language.

2.1.1.- Quantity

The quantity is measured by the number of elements by grammatical category.

2.1.2.- Proportion

The nature of the elements is described globally by the relative proportion of the total number of elements within each category.

2.1.3.- Utility

The usefulness of the vocabulary words selected is measured globally by the frequency and range of these words in free speech.

The values assigned are drawn from "L'Elaboration du Français Fondamental" (Gougenheim, G., et al., Paris, Didier, 1964).

2.2.- Gradation

The order and rate of introduction of each new element may vary from one method to abother. To evaluate the gradation, we measure the intake, density and productivity.

2.2.1.- Intake

The intake is measured for a standard selection of text by the proportion of new elements within the section to the accumulated number of different elements, this for each category of elements.

2.2.2.- Density

The density is defined as the number of new elements per sentence. It is measured globally for the method by the proportion of the sentences with density zero, one, two, three or more.

2.2.3.- Productivity

The productivity is defined here, as the number of possible variant realisations of a structure given the selected number of constituent elements. The structures can be assessed individually and by category at different points within the method. The productivity may be measured as defined, or by some function of the defined variable. We use a logarithmic function in order to reduce the scale of the measurements and to simplify the calculation.

2.3.- Presentation

To measure the means by which a method presents its material, it is necessary to be able to distinguish the different kinds of material within the method. Manual inclusion of coded markers, called "sigla", before the automatic analysis, permits us to measure certain aspects of the presentation.

2.3.1.- Introduction

The number and proportion of the elements are counted according their first occurrence in three kinds of textual material: syntactic presentation, syntactic repetition, or non-syntactic.

2.3.2.- Contextualisation

The amount and proportion of material is measured according to which procedure for contextualising the meaning is favoured. Pictorial contextualisation depends on the different kinds of pictures; differential, on the use of the mother tongue in explanations and translations; and verbal, on the different literary forms of the text.

2.4.- Repetition

The repetition in a method depends on the number of occurences of each element. It may be measured for an element by the total number

of occurences; for a group of elements by the total and by the average repetition per element. The elements are grouped by category, medium, skill and type of exercise.

2.4.1.- Category

The total and average repetition are measured for each category of elements, once for all the original material of the method, and once again for all the original and duplicated material.

2.4.2.- Media

The total word repetition is distributed according to the different media of the method: manual, reader, exercise book, magnetic tape, etc.

2.4.3.- Skill

The total word repetition is distributed according to the skill involved: reading, listening, speaking, writing.

2.4.4.- Type of exercise

The word repetition is measured within each type of exercise: rote, incremental, variational, operational.

3.- Considerations Basic to Data Processing

The adoption of data processing techniques obliged us to organize the project according to the resources and limitations of these techniques.

3.1.- The Equipment

At "Le Centre de traitement de l'Information de l'Université Laval", besides the conventional accounting machinery we had an IBM-1410 computer at our disposal. The installation included a card-reader, six magnetic tape units and an IBM-1403 printer. The memory had a capacity of 60,000 character positions.

3.2.- The Control of the Computer Operations

The instructions according to which the method was manipulated were written in the symbolic programming language Fortran. A number of general operations were already available in the form of sub-programs that could be incorporated in larger instruction sequences (programs) prepared explicity for the project.

In addition, a number of programs furnished with the machine were at our disposition. These programs, henceforth referred to as "control programs" were designed for efficient execution of commonly used operations. In particular, we used control programs to put data from punched cards onto magnetic tape, to print data from magnetic tape onto paper, and to sort data stored on tape.

3.3- The Punching of the Method onto Cards.

All material to be furnished to the computer had to be punched initially onto cards. The punch operator's work is that of a typist; to copy the material as it is presented to him. For this reason, all of the method to be treated had to be in written form.

3.4.- The Editing of the Method

In order to exclude unwanted material and include material not in written form, the method had to be revised manually before being punched onto cards. Unwanted material was barred, and additional material was written in where appropriate. For instance, in order to ensure the identification of pedagogically distinct portions of the text, specially coded "sigla" were inserted in the text.

3.5.- The Programs

The instructions controlling the computer operations on the material of the method had to be formulated in advance. The instructions were grouped into a number of distinct programs rather than in one complete program, for the reasons enumerated:

1. One program would be unwieldy: it would be difficult to prepare, to test, and to modify.

2. One program would be inefficient; in order to use it on the computer with its limited storage capacity, the operations would have to be conceived to economise on storage space rather than on execution time.

3. With several programs, we can put the control programs already available to effective use. This saves programming and testing time for the operations thus implemented.

4. With several programs, it is possible to judge the results at each stage before going on to the next. Modifications and corrections can be incorporated into the succeeding programs.

Each program had to control several types of activity. It had to enter the material of the method (input), work on this material, then store the results (output). The phase where the material is worked on includes all the operations essential to the analysis of the method. Thus the algorithms or logical processes of the analysis are formulated within the programs. First however, they had to be conceived.

Each program must contain all the information upon which its operations are dependant, or have access to such information. That is, the information may be contained in the instructions of the program, or in tables which can be consulted by the instructions. It was decided to include as much as possible of the needed grammatical data in tables, on punched cards, or on tape. In this way, the programs would be independant of the grammar. The latter could then be modified or even replaced by that of another language without necessitating changes in the instructions which use it.

4.- The Preparation of the Method.

The pedagogical material of the method was prepared and subsequently punched onto cards. Both operations were effectuated manually according to prescribed directions.

4.1.- The Pre-Editing of the Method.

Certain parts of the manuals were barred, being non-pertinent to the analysis. This includes:

1. text not in the language being taught,
2. grammatical and lexical lists at the back of the manuals,
3. characters not making up words,
4. text in phonetic script,
5. chapter and exercice numbering,
6. wrong choices in multiple choice exercices,
7. exercice keys at the back of the manuals.

In order to ensure the correct pedagogical ordering of material found in several manuals, the numbering of the pages was modified. Pages of text to be inserted in the principal manual were given the page number of the preceding page, plus a sub-page number from .1 to .9.

Multiple choice and fill-in-the-blanks exercices were completed manually.

To ensure the identification of the pedagogical units of presentation, these were identified by special "sigla". The sigla were coded according to Table 1 and inserted according to the fol-

lowing rules:

1. sigla at the beginning of each pedagogically homogeneous passage,

2. one siglum for each medium of presentation of the passage,

3. one siglum for each series of pedagogically homogeous pictures.

4.2- The Punching of the Method onto Cards.

The text was punched onto cards on a machine with a keyboard of 47 characters. In order to accomodate the letters and punctuation marks encountered in the texts, the following rules were followed:

1. for each alphabetic character, whether small, capital or accented, the unique corresponding keyboard letter,

2. for each accented letter, a number immediately following it according to Table 2.- (a),

3. for each punctuation mark, one or several characters according to Table 2.- (b),

4. for each space, a blank.

The text was punched onto cards according to the following rules:

1. 72 letters and blanks per card in the card colums 1 to 72,

2. continuity from one card to the next in conformity with rules 3 and 4,

3. at least one blank space or punctuation mark between words,

4. no blanks within a word.

The cards were assigned page numbers in the order of presentation of the material. Within each page, the cards were numbered consecutively from 1 up. The page and line numbers occupied the card columns indicated below:

columns 73, 74, 75	page number
column 76	inserted page number
columns 77, 78	line number
column 79	overflow line number.

TABLE 1.- (a) Pictorial Sigla Codes

Type of picture	Column 1	2	3	4	5
with caption	p				
without caption	x				
for distribution		b			
for display		x			
in textbook			t		
in exercise book			e		
in reader			r		
wall picture			h		
picture card			a		
flannelgraph			g		
slides			i		
film strip			x		
motion picture			c		
number (example)				1	2

Table 1.- (b) Textual Sigla Codes

Type of text	Column 1	2	3	4	5
manual	b				
display material	w				
recorded on tape	t				
recorded on disk	d				
recorded on film	f				
orthographic script		o			
phonetic script		p			
recorded, unspaced		r			
recorded, spaced		s			
prose			p		
verse			v		
song with music			m		
dialog			d		
isolated sentence			s		
isolated phrases			f		
isolated words			w		
caption or title			y		
for reading				r	
for listening				l	
explanation, differential				x	
explanation, vernacular				z	

TABLE 1.- (b) Textual Sigla Codes

Type of text	Column 1	2	3	4	5
imitation, copying				c	
dictation, transcription				d	
for reading aloud				e	
incremental imitation				i	
completion				f	
multiple choice				s	
alteration				a	
paraphrasing, rephrasing				p	
question and answer				q	
oral composition				o	
written composition				w	
translation from vernacular				t	
translation into vernacular				v	
grammar				g	
lexical list				y	
syntactic presentation					t
syntactic repetition					r
non-syntactic					x
omitted from cards					o

TABLE 2.- Diacritic Transcription

Text character	Punch character
(a) Accents	
ç (cedilla)	Ignored
/ (acute)	1
^ (circumflex)	2
` (grave)	3
¨ (diaeresis)	4
(b) Punctuation	
.	.
,	,
-	-
(	(
)	)
—	—
?	$
!	- .
:	..
;	.,
¨,«	((
¨,»	))
...	...

5.- The Analysis

The organization of the analysis is outlined graphically in Figure 1. The series of programs includes the following basic steps:

1. the identification of the graphic elements: words, punctuation, sigla,
2. the identification of the phrase structures,
3. the identification of the clause and sentence structures,
4. the counting of the elements identified according to their different classes,
5. computation of the measurements dependent upon the elements and their counts, and
6. the presentation of the results in lists, tables, and graphs.

5.1- Cards to Tape

The material of the method was transfered from punched cards to magnetic tape with the IBM-1410 and the control program "Cards to Tape". The contents of twenty cards was grouped within each tape record. The storage tape was labelled as "Text".

5.2- Program P1: Word Separation

The words contained in the tape records of "Text" were separated into individual records. Punctuation and sigla were treated as separate words.

Each record of one word was assigned the page and line number of the corresponding card.

An order number was assigned to each record. This number corresponded to the consecutive integer for each word, with an increment of 0.2 for each punctuation mark.

Each record was assigned a duplication number corresponding to the number of sigla identifying the passage of the text.

Each record was assigned a presentation number based on the fifth position of the sigla code. This distinguishes text for presentation, text for repetition, and material not in sentence form.

The new records containing sigla were stored consecutively on the tape labelled "Text Sigla". Each record was punched onto a card for permanent storage.

The records containing words and punctuation were put onto the tape labelled "Textual Words".

5.3- Program P 2: Word Identification

The records of "Textual Words" were sorted in alphabetical order with the control program "Sort and Merge". The resulting tape was called "Alphabetic Words".

The words and punctuation were identified by comparaison with the prepared lists "Vocabulary Words" and "Functional Words".

The information contained in the input records was transferred to the output records.

In addition, certain data contained in the lists was transferred to the identified elements. This includes a grammatical category number, a word identity number, and, for certain words with multiple functions, an ambiguity number.

For vocabulary words with regular grammatical endings, a new record containing the ending was produced. The ending was identified

in the list "Grammatical Endings" and assigned the pertinent data as for the words. For each different vocabulary word the corresponding utility parameters of the "Vocabulary List" were accumulated by word category. The totals were assigned to a record with fictive identity and order numbers to distinguish them from the words.

The records containing words, punctuation, endings and utility counts were put onto the tape "Alphabetic Identities".

The first occurences of unidentified elements were punched onto the cards "Unidentified Words" with their input information.

A list of each different type of input element with the output data of its first occurrence and a count of the number of its occurrences was printed under the title "Identified Words".

5.4.- Correction of Word Identities

The words in the card deck "Unidentified Words" were identified manually; that is to say, they were assigned numbers analogous to those of the identified words. The identifying numbers were punched onto the cards now called "Word Corrections".

The list "Identified Words" was revised manually to ensure their identification and their spelling. All modifications were punched onto cards and included in the deck "Word Corrections".

5.5.- Program P 3: Word Correction

The punctuation, words, and grammatical endings of the card deck "Word Corrections" were compared sequentially with those contained on the tape "Alphabetic Identities". The corresponding items were modified, with extra items being added as indicated on the cards. The output tape was called "Correct Identities."

5.6.- Program P 3.5: Concordance Counts

The records of "Correct Identities" were sorted by the control program "Sort and Merge" according to the hierarchy: category number, alphabetic order, textual order number. From the resulting concordance, on the tape labelled "Concorded Identities" was selected the first record for each different type of punctuation, graphic word, and grammatical ending. The number of tokens of each type of element was counted to be included in the output record on the tape "Identity Types".

A list, printed directly from the tape, included the following information for each item: category number, identity number, page and line number of the first occurrence, order number, and token count. The latter gives the amount of repetition for each type. The list kept the same order as the tape, that is, first by category, then by alphabetic order.

5.7.- Program P 4: Phrase Identification

The records of "Correct Identities" were sorted in textual order onto the tape "Textual Identities". Next, in one pass over the identified tokens in their textual order, we resolved the functional ambiguities of word identities, counted the types and tokens of the words and grammatical endings, identified the phrase structures, and prepared the sequence of phrases found in each sentence.

The functional ambiguities of certain problem words were specified by the ambiguity number at the moment the word was identified (P 2). This number refers us to a sequence of rules contained in the card deck "Ambiguity Rules". The rules give the dependence of the category and identity numbers on the neighbouring elements, Both numbers are modified as indicated by the rules.

The number of types and tokens are counted by category of elements for each block of 500 textual words at each blocking point. The counts were put onto a record of the tape "Identity Counts".

The distribution of the types of element in the whole method was counted according to their introduction. That is, the number of first occurrences of vocabulary words, functional words, and grammatical endings was counted according to the presentation number assigned at word separation (P 1). The counts were included in the final record of "Identity Counts".

The utility counts accumulated during the word identification (P 2) were transferred to the final record of "Identity Counts".

The phrases were identified by comparing maximal sequences of word and punctuation categories in textual order with the phrase structures stored on the cards "Phrase Dictionary". The identified phrases were assigned category and identity numbers, and the page, line, duplication numbers of the terminal punctuation of the sentence. The order number of the terminal punctuation was given the increment 0.3 for the output record of each phrase. The output tape was called "Phrase Structures."

The category numbers of the identified phrases were put into sequence for each sentence. Each phrase sequence was assigned the same information as the phrases making it up, less the category and identity number. The output records were put onto the tape "Textual Sequences".

For each sentence we counted the number of new types of vocabulary word, functional word, and grammatical ending. The three counts were included in the records of "Textual Sequences."

5.8.- Program P 5: Clause Identification

The phrase sequences were broken down into clause structures which were in turn identified and grouped into sentence structures. The algorithm for cutting the phrase sequences called upon rules contained in the card deck "Clause Rules" to determine which of the phrase sequence elements delimited different clauses of the sentence.

The cutting algorithm had to be implemented only once for each series of identical sequences, the resulting clauses being replicated for the appropriate number of sequences. Hence it was economical to first sort "Textual Sequences" in order of the sequences' alphabetic value onto a new tape "Alphabetic Sequences".

Each clause structure was assigned a category number depending on its first element. The sequence of clause category numbers of a sentence gave the sentence structure, which was assigned a category number depending on the number of clauses it contained. The input information with the phrase sequences was transferred to the clause and sentence structure output records. For the clauses, the order number was given an increment of 0.4; for the sentences, 0.5. The output records were put on the tape "Clauses and Sentences".

5.9.- Program P 6: Structure Concordance

For the input, the tapes "Phrase Structures" and Clauses and Sentences" were sorted according to category number, alphabetic value of the token, and order number onto the tape "Categorized Structures". Each different clause and sentence structure was assigned an identity number, and put onto the tape "Concorded Structures".

The first token of each structure was put onto the tape "Structure Types". Included in each output record was the information of the first token plus the number of tokens of the structure type. A printed

list of the output tape gave us the structure types in order of category and alphabetic value of the structure. The information displayed included the category and identity numbers, the page and line numbers of the first occurrence, the order number and the token count.

5.10.- Program P 7: Structure Counts

The structures of "Structure Types" were sorted in textual order onto the tape "Textual Structures." They were counted in this order for the number of types and tokens within each category for each section of text of 500 - word occurrences.

Each new structure type was counted according to the presentation number of its first occurrence. This introduction count was done for the whole manual.

The number of new phrase, clause and sentence structure types was counted by sentence. Each sentence was then counted according to its number of new vocabulary words, functional words, grammatical endings, and structures. These density counts were accumulated for the whole method.

At each 500 - word occurrence blocking point, the corresponding record of "Identity Counts" was read in. The type-token structure counts and the productivity counts were combined with the input counts for the words and grammatical endings for the output records of "Counts".

The productivity of the structures was calculated in terms of their immediate constituents. The logarithm of the type counts for each element of each structure was accumulated for the phrases, for the clauses, and for the sentences.

The introduction and density counts were included in the final record of "Counts".

5.11.- Program R I: Block Results

The data contained in "Counts" was manipulated to give part of the measurements defined in chapter II. The results were printed out in tables. The headings and other worded indications in the table were read in from the prepared card deck "Block Titles".

Included in the results for each block were the intake and productivity of the gradation (ref. 2.2.) . For the method as a whole, results were produced for the quantity, proportion, and utility of the selection (ref. 2.1.), the density of the gradation (ref. 2.2.2.), the introduction of the presentation (ref. 2.3.1.), and the repetition by category (ref. 2.4.1.).

5.12.- Program R 2 (a & b) : Sigla Analysis

The list of sigla printed out after the word separation (P1) was verified manually. Any errors or modifications were taken up in the card deck "Text Sigla". The corrected deck served as data for tabulations

of presentation and repetition. The number of word occurrences governed by each sigla was given by the difference between order numbers between sigla. The occurrences were accumulated according to each of the coded characters in the sigla. The totals were stored on the tape "Sigla Tables".

The final results of the sigla analysis were organized and printed out in tables by the program R 2 (b). The card deck "Sigla Titles" controlled the organization of the tables and supplied the headings as well.

The results included the contextualization of the presentation (ref. 2.3.2.), and the repetition by media, shill, and type of exercise (ref. 2.4.).

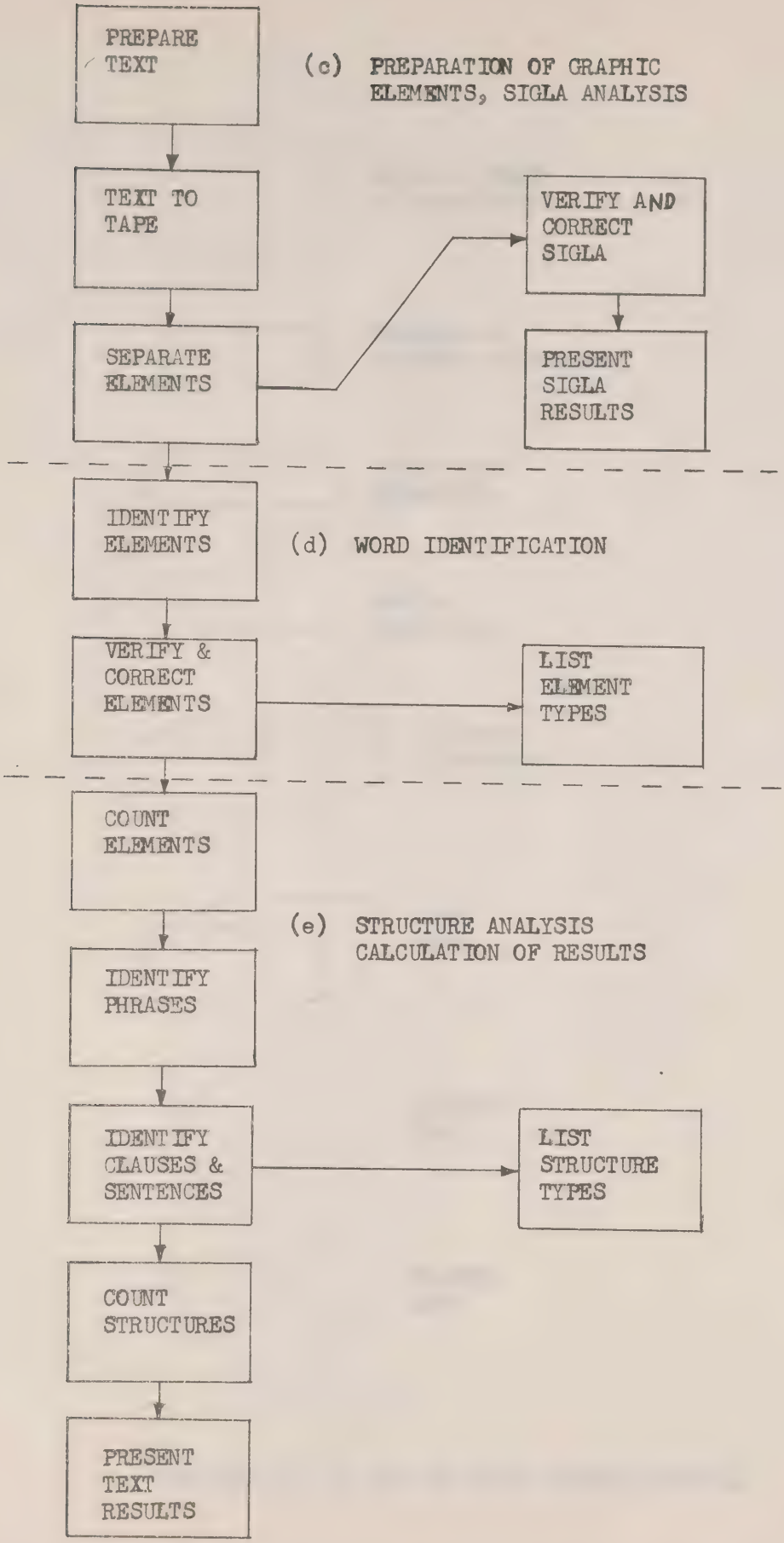


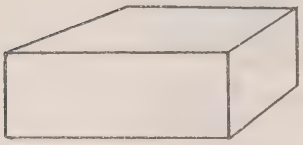
FIGURE 1.- (a) BLOCK DIAGRAM OF THE ANALYSIS

LOGICAL GROUP
OF OPERATIONS

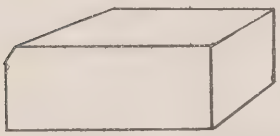
PROGRAM OF
MACHINE OPERATIONS

CARD PUNCH
OPERATION

MANUAL
OPERATION

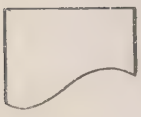


PRINTED
DOCUMENTS



CARD
DECK

MAGNETIC
TAPE



PRINTED
LIST

FIGURE 1.- (b) KEY TO BLOCK DIAGRAM SYMBOLS

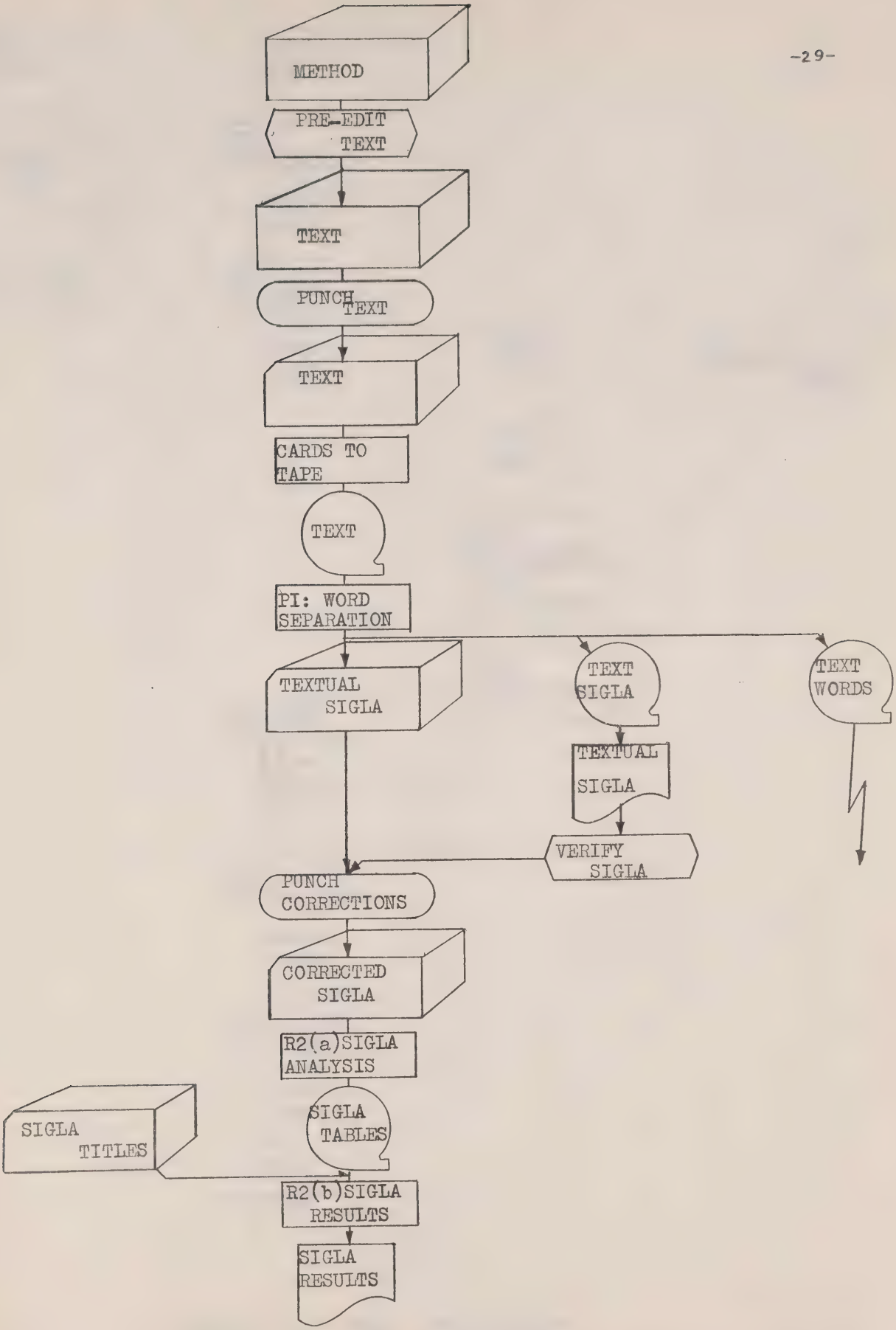


FIGURE 1.- (c) PREPARATION OF GRAPHIC ELEMENTS, SIGLA ANALYSIS

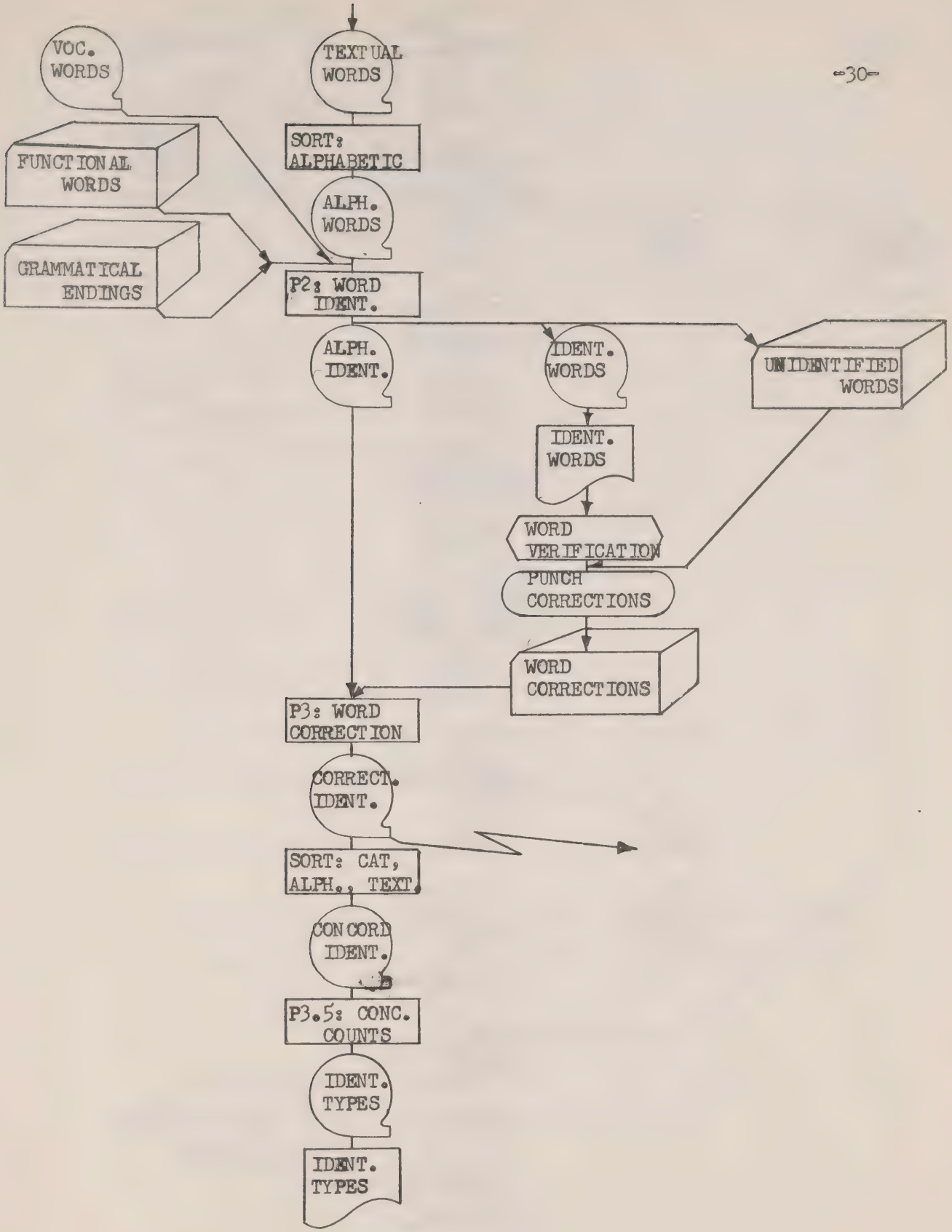


FIGURE 1.- (d) WORD IDENTIFICATION

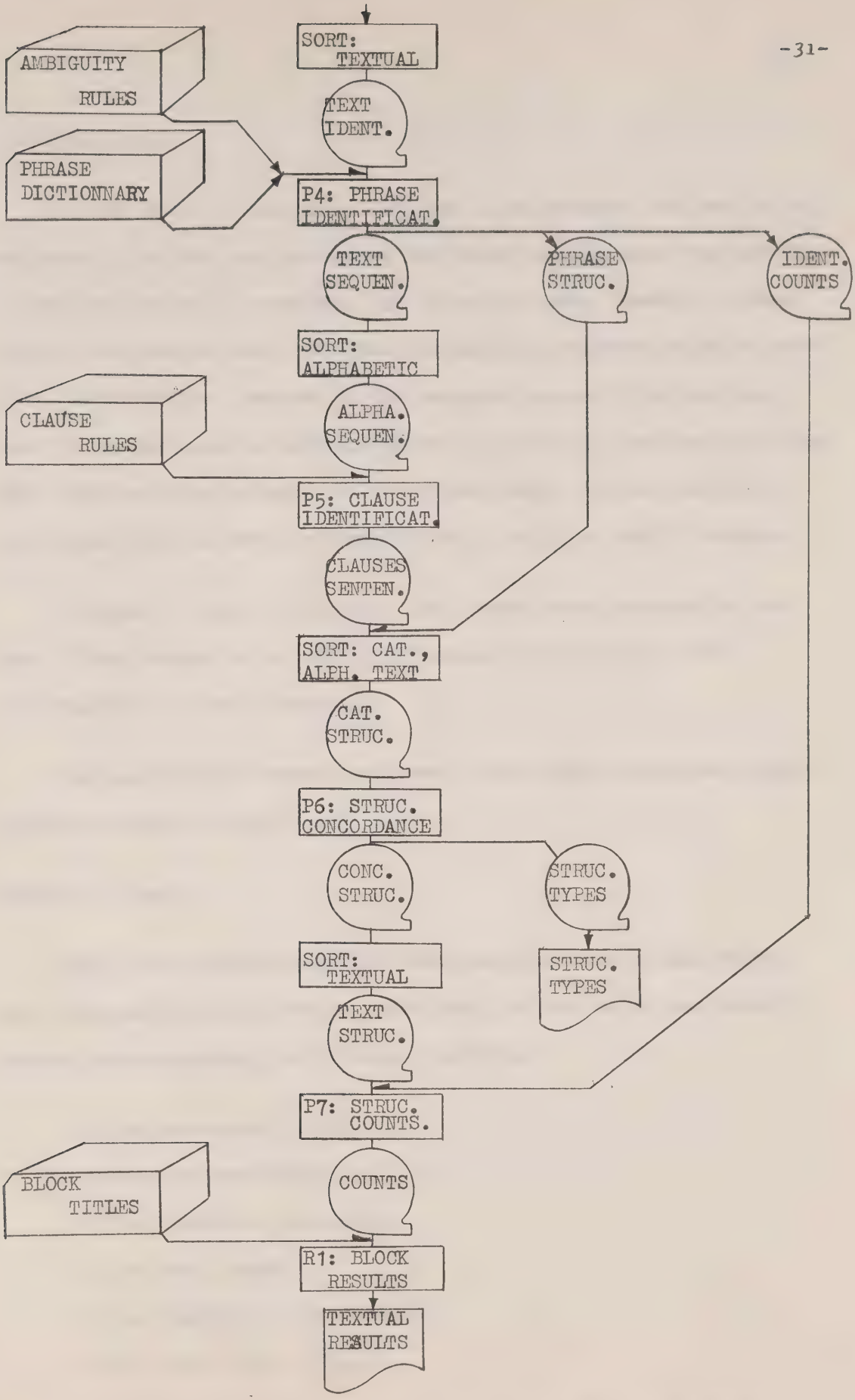


FIGURE 1.-(e) STRUCTURE IDENTIFICATION, CALCULATION OF RESULTS.

6.- The Input Data

A great deal of linguistic data had to be prepared for use in the analysis. The data was formulated for compatibility with the previously outlined analytic procedure. An effort was made, however, to keep the two distinct so that both the data and the procedure could be modified independently. Because of the original nature of the project, much of the work was tentative. We anticipated the lengthening of the data lists and the inclusion of additional data. All but the most voluminous data was kept on cards where it could be readily changed.

A number of output procedures and formats were governed by card data. This allowed us to print the results in any chosen order with English or French headings.

The coordination between programs, input data, and output results is demonstrated in Figure 1.

6.1.- Vocabulary Words

This list contains about 800 vocabulary words from among those most frequently used in French. The data for each entry consisted of numbers characterizing the following entities:

1.- the grammatical category,

2.- the word root identity,

3.- the functional ambiguity,

4.- the root length,

5.- the frequency (ref. 2.1.3.),

6.- the range (ref. 2.1.3.).

Because of its length and the stability of its information, the list was put on magnetic tape.

The tape records were sorted in alphabetical order in preparation for use by the program P 2.

Basically, only one form of each word was included in the list. For certain words, however, the inflected forms corresponded to alphabetical positions differing widely from the listed form. To simplify the search procedures of program P 2, extra forms were included, especially for the verbs, where the separation exceeded three list positions.

6.2.- Grammatical Endings

About 80 of the most regular endings were stored on cards. Each was assigned numbers to characterize its identity and its grammatical category. They were listed in alphabetical order by category.

6.3.- Functional Words

This list contains the functional words, the punctuation marks, and any vocabulary words not in "Vocabulary Words" but foreseen as needed (e.g. proper nouns).

The words were put punched cards with identity and category numbers. The prepared deck was in alphabetic order.

There are several reasons why this list was kept separate from "Vocabulary Words":

1 - A simpler and faster search procedure is possible for elements with no flectional variants.

2 - The number of elements is relatively limited: about 300 functional word forms and about 15 combinations of punctuation marks.

3 - The classification of the functional words presents special problems. The categories needed for the analysis depend largely on the analysis itself. Functional ambiguities are numerous.

4 - The combinations of basic punctuation marks encountered varies from one method to another. Their utilization, hence their grammatical function, varies as well.

5 - Extra words can be added at will.

6.4.- Ambiguity Rules

For each word type characterized by an ambiguity number corresponds a sequence of tests and operations encoded on the punched cards. The rules define the correspondance between the textual environment of an ambiguous element and the changes to be made in the identification of the elements. For each ambiguity number there is a card containing one rule. Each rule may call upon another, this building up sequences of tests that can explore beyond the elements immediately before and after the one in question.

Each rule contains up to six numbers which occupy specified positions in the card:

1 - The ambiguity number characterizes the rule and specifies whether the operation applies to the element in question, the preceding element, or the following element.

2 - The operator number specifies the operation:

a) test for equality of the operator number with the tested elements category number,

b) test for equality of the operator number with the tested element's identity number,

c) eliminate the tested element,

d) replace the tested element,

e) insert an element after the tested element.

3 - Three numbers (3rd, 4th, 5th) give the new category, identity, and ambiguity numbers to be assigned by the operation, if appropriate. The two test operations (a & b) call for these numbers only if the test is positive.

4 - The sixth number is the ambiguity number of the next rule to be invoked. It does not apply after a positive test.

About 100 rules were included for the first analysis. They were ordered by increasing ambiguity number.

6.6.- The Phrase Dictionary

About 250 phrase structures that could be encountered in the grammatical hierarchy were prepared in advance. To each was assigned an identity number. The phrases were grouped into categories according to their function in the clause and given an appropriate category number. They were then put on cards and ordered by increasing alphabetical value of the structure.

6.7.- The Clause Rules

The rules for dividing a sentence containing two or more conjugated verbs into its constituent clauses were composed in terms of certain critical clause elements. In the program P 5 the phrase sequence elements corresponding to possible conjunctive elements are arranged in a skeletal sequence. The list of clause rules contains about 50 possible skeletal sequences, representing possible phrase sequences up to the second verb. Four numbers delimit a clause within the phrase sequence, thus permitting the abstraction of the clause from the phrase sequence. A fifth number gives the category of the clause. The procedure may be iterated for sentences with more than two clauses.

Two of the numbers mentioned give, respectively, the two limiting elements of the clause within the skeletal sequence. The other two numbers determine whether or not the limiting elements are included or excluded from the part of the phrase sequence making up the clause.

6.8.- Word Corrections

One card was prepared for each element type to be modified in P 3. The cards contained the following information:

1 - the uncorrected spelling of the element,

2 - the corrected spelling of the word, where appropriate,

3 - the corrected identity, category, and ambiguity numbers where appropriate,

4 - an operator number to determine one of three operations on the elements to be corrected:

a) modification of the element as indicated,

b) elimination of the element,

c) insertion of the indicated number of extra elements.

The operation specified on the word correction card was carried out for each of the corresponding elements of the uncorrected tape "Alphabetic Words".

The word corrections were sorted alphabetically to correspond to the order of "Alphabetic Words". For each correction calling for insertions, the elements to be inserted were included, one per card, immediately after the correction card.

6.9.- Block Titles

The headings chosen for the output result tables were put onto cards in the order and position needed for the output.

6.10.- Corrected Sigla

Each verified sigla card contained a siglum as defined (ref. 4.1. and Table 1), and the order number of the word preceding it in the text. The sigla cards were sorted in textual order.

6.11.- Sigla Titles

This card deck was prepared to control the output result tables of

program R 2 (b). Each card contained four possible kinds of information:

1 - a number indicating the sub-series of program instructions appropriate for the calculation of the desired measurement,

2 - numbers indicating the data of "Sigla Tables" entering into the calculations,

3 - the heading desired for the line of results to be printed,

4 - a number controlling the line and page spacing of the printed results.

7.- Problems Encountered During Trials

The discussion of the analysis and of the input data assumes the workability of the system. In fact, it treats the system in its ultimate state, after its trial and succeeding modification. A number of problems arising during the trials retarded the realization of the project.

7.1.- Programming Problems.

The programs were first prepared for use with a machine memory of 40,000 character positions. Practically, this relatively low limit meant breaking up the logical sequence of machine operations into an awkward series of programs. That is to say, it took more than the ideal number of passes on the machine to handle the input material. It meant as well that the linguistic data had to be organized into dense tables and entered in small portions into the memory. As a result the first trial production was extremely uneconomical; it required two and a half times the machine time required by later productions.

The problem of machine memory was solved by the addition of another memory unit of 20,000 character positions. To capitalize on the increased resources of the computer, the programs had to be re-organized. Parts of the programs were rewritten, others omitted. The transformed programs had to be retested, and a second trial run produced.

7.2.- Program Testing

The functioning of each program was verified in a series of tests on the computer. After each testing run, the results were examined for errors due to faulty instructions. The program was subsequently modified then

retested. This "debugging", as it is called, is standard procedure.

This stage in the preparation of the programs proved far more lengthy than anticipated. First of all, the complexity of the instructions necessary for non-numeric data handling served to increase the incidence of error and the difficulty of correction. For instance a faulty instruction may cause the obliteration of another instruction well removed from it; when invoked, the missing instruction will cause a program halt, but will not reveal the source of the error. Only by lengthy and tedious examination of the program instructions and of the test results could certain errors be traced. Had the programs been simpler, a short inspection would have sufficed.

The sharing of machine time imposed serious restrictions on the rhythm of program testing. A minimum lapse of one day could be expected between the submission of a program for testing and the reception of the test results. Counting correction time, a series of ten or twelve tests could amount to a month for a single program. As much as possible, the programs were tested concurrently, but where the output of a program was needed as input to the succeeding one, the latter could only be tested consecutively.

During the testing of the programs, the conditions of machine utilization were never near the level achieved during the period of production. First, there was only one operator available to run the computer. During periods of heavy demand for machine time, it could take up to a week for a job to be treated. When inevitable operator absences or machine malfunctioning caused the loss of a day, it could be several extra days before the accumulated work was done.

The programs that were modified when the extra memory unit arrived had to be retested. Inevitably, the new unit temporarily impeded normal machine usage while its own defects were being tested and eliminated. A month after the installation, several of the programs could still not be used due to errors generated by the new equipment.

7.3.- The Trial Run

With the programs finally working, the system had to be subjected to a test case in order to judge the adequacy of the input linguistic data and the appropriateness of the program logic. The trial run exposed several problems that had to be solved before serious production could start.

On the basis of the trial results, the procedure for phrase and clause identification was modified to decrease production time and improve the results. It also turned out that the input data for phrase identification was not sufficiently extensive to handle the phrases encountered. The data file was enlarged upon accordingly.

Some of the measurements to be tabulated in the results were tentative. In particular, the figures for the productivity (ref. 2.2.3.) moved us to replace its direct measurement with the logarithmic function of its value. Other minor changes in the measurements and their representation were incorporated into the programs.

8.- Interpretation of the Results

The immediate results as put out by the computer may be on punched cards or magnetic tape rather than on printed pages. They are transferred to paper wherever readable lists are desired. Hence all the results are treated as if in readable form.

8.1.- The Working Lists

After each step of the analysis any or all of the output data may be considered as intermediate results. As indicated in Figure 1, the working lists "Word Corrections", "Identified Words", and "Textual Sigla" are indispensable for the correction of the data at crucial points in the analysis. Their use is discussed in Chapter 5.

A full concordance of the words, punctuation, word endings and structures of each method exists in the lists "Concorded Identities" and "Concorded Structures". These two lists are useful for detailed examination of the method at the level of its grammatical elements, but are extremely voluminous.

8.2.- The Type Lists

The lists "Identity Types" and "Structure Types" summarize the concordances of the words, punctuation, word endings and structures. Each graphically different element is listed with its identifying information and individual token count for the whole method. A copy of the lists is included in Appendix I (1) with a key to the information contained in them.

(1) Appendix 1 consists of machine print-out; hence, it exists only one copy.

8.3.- The Block Results

The intake and the productivity were calculated at 500-word intervals throughout the text. The results are tabulated in "Textual Results". A sample list is presented in Appendix I .

8.3.1.- The Productivity

The productivity degree is the logarithm to the base 10 of the productivity. In Table 3 the total degree is the sum of the productivity degrees of the structures, and the average degree is the arithmetic mean of the total.

For example, the value 5.1 for the average degree of the phrases in Table 6.- (a) may be interpreted as meaning a combinatorial productivity of about 100,000 ($10^{5.1}$) for each phrase, or of 17,500,000 ($176 \times 10^{5.1}$) for all the phrases of the method.

8.3.2.- The Intake

The intake was calculated by dividing the number of new types within a given category by the number of its tokens in the 500-word block. Thus it is not strictly proportional to the amount of new material. High values may be due to a large number of new types or a low number of repetitions. The ambivalence was eliminated from the graphic representation (ref. 8.5.).

8.4.- The Global Results

All of the measurements, including the intake and the productivity, were tabulated for the method globally in "Sigla Results" and "Textual Results". They appear in Tables 3 to 11 for the two methods analysed.

A number of the measurements depend on the analysis of the coded sigla. The process of pre-editing proved to be somewhat unsatisfactory. The diversity of the manual operations demanded at the pre-editing stage complicated the task for the recent initiates doing it. The multiplicity of possible sigla combinations contributed to the difficulty. Because of the tentative nature of the first analyses, the pre-editing was verified only once. The obvious coding errors were taken up at the sigla correction stage, the rest remained.

In Table 8.- (b) a systematic error of interpretation caused most of the verbal contextualization to be counted as "list" rather than as "prose" or "dialog".

In Tables 8, 9, and 11 the results must be considered approximative. The multiple repetition in Table 10 is approximative as well.

8.5.- The Graphic Results

Figures 2 to 10 are graphic representations of the results in Tables 3 to 11 and of the block results (ref. 8.3.). The following indications eliminate possible ambiguities of interpretation.

1. In Figure 3, the values of the range were multiplied by a factor of 3 to enhance their presentation.

2. The number of new types rather than the type-token ratio was chosen to represent the intake in Figure 4. This simplifies the interpretation of the graph: where high, there are many new types; where low, few.

3. In Figure 4, the new types were counted within distinct, consecutive, 1000-word segments of running text.

4. The height of the columns in Figure 6 represents the proportion of the sentences presenting one or more new elements.

5. Being based on the approximate results of the corresponding tables, the Figures 7, 8 and 10 are approximative. The multiple repetition of Figure 9 is as well.

6. The multiple repetition of Figure 9 represents the extra repetition due to textual material duplicated in media other than the ones punched onto cards.

TABLE 3.- (a)

VOIX ET IMAGES DE FRANCE — SELECTION, QUANTITY

CATEGORY	TYPES (NUMBER)
NOUNS	814
VERBS	273
ADJECTIVES	147
ADVERBS	17
TOTAL, VOCABULARY	1254
TOTAL, GRAMMAR WORDS	183
TOTAL, WORD ENDINGS	57
PHRASES	176
CLAUSES	1641
SENTENCES	38
TOTAL, STRUCTURES	1855
TOTAL, GLOBAL	3349

Table 3.- (b)

FRENCH THROUGH PICTURES — SELECTION, QUANTITY

CATEGORY	TYPES (NUMBER)
NOUNS	445
VERBS	99
ADJECTIVES	107
ADVERBS	3
TOTAL, VOCABULARY	654
TOTAL, GRAMMAR WORDS	140
TOTAL, WORD ENDINGS	46
PHRASES	138
CLAUSES	1070
SENTENCES	34
TOTAL, STRUCTURES	1242
TOTAL, GLOBAL	2082

TABLE 4.- (a)

VOIX ET IMAGES DE FRANCE — SELECTION, PROPORTION AND UTILITY

CATEGORY	TYPES (NUMBER)	PROPORTION (PERCENT)	FREQUENCY (AVERAGE)	RANGE (AVERAGE)
NOUNS	814	64.9	24.7	9.4
VERBS	273	21.7	108.6	21.8
ADJECTIVES	147	11.7	40.7	15.1
ADVERBS	17	1.3	75.2	32.1
TOTAL, VOCABULARY	1254	100.0	45.5	13.0

TABLE 4.- (b)

FRENCH THROUGH PICTURES — SELECTION, PROPORTION AND UTILITY

CATEGORY	TYPES (NUMBER)	PROPORTION (PERCENT)	FREQUENCY (AVERAGE)	RANGE (AVERAGE)
NOUNS	445	68.0	35.0	12.8
VERBS	99	15.1	239.2	33.8
ADJECTIVES	107	16.3	44.8	13.6
ADVERBS	3	.4	44.0	29.3
TOTAL, VOCABULARY	654	100.0	67.5	16.2

TABLE 5.- (a)

VOIX ET IMAGES DE FRANCE — INTAKE, GLOBAL

CATEGORY	ELEMENTS (NUMBER)	TOKENS (NUMBER)	INTAKE (PERCENT)
NOUNS	814	7415	10.9
VERBS	273	4069	6.7
ADJECTIVES	147	1313	11.1
ADVERBS	17	129	13.1
TOTAL, VOCABULARY	1254	12932	9.6
TOTAL, GRAMMAR WORDS	183	21183	.8
TOTAL, WORD ENDINGS	57	3720	1.5
PHRASES	176	22072	.7
CLAUSES	1641	6288	26.0
SENTENCES	38	5457	.6
TOTAL, STRUCTURES	1855	33817	5.4
TOTAL, GLOBAL	3349	71652	4.6

TABLE 5.- (b)

FRENCH THROUGH PICTURES — INTAKE, GLOBAL

CATEGORY	TYPES (NUMBER)	TOKENS (NUMBER)	INTAKE (PERCENT)
NOUNS	445	6719	6.6
VERBS	99	1926	5.1
ADJECTIVES	107	1649	6.4
ADVERBS	3	25	12.0
TOTAL, VOCABULARY	654	10319	6.3
TOTAL, GRAMMAR WORDS	140	16443	.8
TOTAL, ENDINGS	46	3120	1.4
PHRASES	138	16607	.8
CLAUSES	1070	4652	23.0
SENTENCES	34	3988	.8
TOTAL, STRUCTURES	1242	25247	4.9
TOTAL, GLOBAL	2082	55129	3.7

TABLE 6.- (a)

VOIX ET IMAGES DE FRANCE — PRODUCTIVITY, GLOBAL

CATEGORY	ELEMENTS	TOTAL DEGREE	AVERAGE DEGREE
PHRASES	176	913	5.1
CLAUSES	1641	20980	12.7
SENTENCES	38	656	17.2
TOTAL, STRUCTURES	1855	22549	12.1

TABLE 6.- (b)

FRENCH THROUGH PICTURES — PRODUCTIVITY, GLOBAL

CATEGORY	TYPES	TOTAL INDEX	AVERAGE INDEX
PHRASES	138	630.	4.5
CLAUSES	1070	14047.	13.1
SENTENCES	34	479.	14.1
TOTAL, STRUCTURES	1242	15157.	12.2

TABLE 7.- (a)

VOIX ET IMAGES DE FRANCE — GRADATION, DENSITY

CATEGORY	DEGREE 0	1	2	3	MORE
	(PERCENT OF ALL SENTENCES)				
TOTAL, VOCABULARY	82.5	13.7	2.6	.7	.2
TOTAL, GRAMMAR WORDS	97.0	2.7	.2	.0	.0
TOTAL, WORD ENDINGS	99.0	.9	.0	.0	.0
TOTAL, STRUCTURES	70.8	25.6	2.6	.5	.2
TOTAL, GLOBAL	60.7	26.5	7.9	2.5	2.1

TABLE 7.- (b)

FRENCH THROUGH PICTURES — GRADATION, DENSITY

CATEGORY	DEGREE 0	1	2	3	MORE
	(PERCENT OF ALL SENTENCES)				
TOTAL, VOCABULARY	85.5	13.2	.9	.2	.1
TOTAL, GRAMMAR WORDS	96.7	3.0	.2	.0	.0
TOTAL, WORD ENDINGS	98.8	1.1	.0	.0	.0
TOTAL, STRUCTURES	73.8	22.1	3.1	.6	.2
TOTAL, GLOBAL	62.1	28.2	6.8	1.7	1.0

TABLE 8.- (a)

VOIX ET IMAGES DE FRANCE — PRESENTATION

PICTURE CONTEXT	NUMBER OF UNITS	PERCENT OF TOTAL	PER THOUSAND TEXT WORDS
TOTAL PICTURES	2194	99.9	63.0
WITH LEGEND	2194	99.9	63.0
WITHOUT LEGEND	0	.0	.0
FOR DISTRIBUTION	0	.0	.0
IN MANUALS	0	.0	.0
FOR DISPLAY	2099	95.6	60.2
FOR FIXED PROJECTION	2099	95.6	60.2
AS MINUTES OF FILM	0	.0	.0
DIFFERENTIAL CONTEXT			
TRANSLATION	0	.0	.0
EXPLANATION	0	.0	.0
INSTRUCTIONS	0	.0	.0
VERBAL CONTEXT			
TOTAL TEXT WORDS	34820	100.0	
DIALOG	22998	66.1	
PROSE	3862	11.1	
SONG AND VERSE	0	.0	
LIST	7930	22.7	

TABLE 8.- (b)

FRENCH THROUGH PICTURES — PRESENTATION			
PICTURE CONTEXT	NUMBER OF UNITS	PERCENT OF TOTAL	PER THOUSAND TEXT WORDS
TOTAL PICTURES	1676	99.9	61.4
WITH LEGEND	1171	69.8	42.9
WITHOUT LEGEND	505	30.1	18.5
FOR DISTRIBUTION	0	.0	.0
IN MANUALS	668	39.8	24.4
FOR DISPLAY	1862	111.0	68.2
FOR FIXED PROJECTION	1004	59.9	36.7
AS MINUTES OF FILM	0	.0	.0
DIFFERENTIAL CONTEXT			
TRANSLATION	94	99.9	3.4
EXPLANATION	94	99.9	3.4
INSTRUCTIONS	0	.0	.0
VERBAL CONTEXT			
TOTAL TEXT WORDS	27289	101.4	
DIALOG	11	.0	
PROSE	975	3.6	
SONG AND VERSE	25	.0	
LIST	25878	96.2	

TABLE 9.- (a)

VOIX ET IMAGES DE FRANCE — PRESENTATION, INTRODUCTION

CATEGORY	PRESENTATION (PERCENT	REPETITION OF ALL NEW	NON-SYNTACTIC TYPES)
TOTAL, VOCABULARY	58.2	39.4	2.3
TOTAL, GRAMMAR WORDS	74.3	21.8	3.8
TOTAL, WORD ENDINGS	59.6	38.5	1.7
TOTAL, STRUCTURES	42.6	57.3	.0
TOTAL, GLOBAL	50.4	48.4	1.1

TABLE 9.- (b)

FRENCH THROUGH PICTURES — PRESENTATION, INTRODUCTION

CATEGORY	PRESENTATION	REPETITION	NON-SYNTACTIC
	(PERCENT OF ALL NEW TYPES)		
TOTAL, VOCABULARY	60.5	39.4	.0
TOTAL, GRAMMAR WORDS	62.8	37.1	.0
TOTAL, WORD ENDINGS	41.3	58.6	.0
TOTAL, STRUCTURES	38.4	61.5	.0
TOTAL, GLOBAL	47.1	52.8	.0

TABLE 10.- (a)

VOIX ET IMAGES DE FRANCE — REPETITION BY CATEGORY

CATEGORY	TYPES (NUMBER)	SIMPLE TOKENS (NUMBER)	SIMPLE REP. (AVE.)	MULTIPLE TOKENS (NUMBER)	MULTIPLE REP. (AVE.)
NOUNS	814	7415	9.1		
VERBS	273	4069	14.9		
ADJECTIVES	147	1313	8.9		
ADVERBS	17	129	7.5		
TOTAL, VOCABULARY	1254	12932	10.3	17506	13.9
TOTAL, GRAMMAR WORDS	183	21183	115.7	29329	160.2
TOTAL, WORD ENDINGS	57	3720	65.2	5030	88.2
PHRASES	176	22072	125.4	30333	172.3
CLAUSES	1641	6288	3.8	8639	5.2
SENTENCES	38	5457	143.6	7660	201.5
TOTAL, STRUCTURES	1855	33817	18.2	46632	25.1
TOTAL, GLOBAL	3349	71652	21.3	98497	29.4

TABLE 10. - (b)

FRENCH THROUGH PICTURES — REPETITION BY CATEGORY

CATEGORY	TYPES (NUMBER)	SIMPLE TOKENS (NUMBER)	SIMPLE REP. (AVE.)	MULTIPLE TOKENS (NUMBER)	MULTIPLE REP. (AVE.)
NOUNS	445	6719	15.0		
VERBS	99	1926	19.4		
ADJECTIVES	107	1649	15.4		
ADVERBS	3	25	8.3		
TOTAL, VOCABULARY	654	10319	15.7	18299	27.9
TOTAL, GRAMMAR WORDS	140	16443	117.4	29127	208.0
TOTAL, WORD ENDINGS	46	3120	67.8	5502	119.6
PHRASES	138	16607	120.3	29465	213.5
CLAUSES	1070	4652	4.3	8300	7.7
SENTENCES	34	3988	117.2	7224	212.4
TOTAL, STRUCTURES	1242	25247	20.3	44989	36.2
TOTAL, GLOBAL	2082	55129	26.4	97917	47.0

TABLE 11.- (a)

VOIX ET IMAGES DE FRANCE — REPETITION, DISTRIBUTION

ACCORDING TO SKILL	NUMBER OF TOKENS	PERCENT OF TOTAL	PER HUNDRED TEXT WORDS
TOTAL WORDS OF METHOD	48112	99.9	138.1
LISTENING	21591	44.8	62.0
READING	33199	69.0	95.3
SPEAKING	18495	38.4	53.1
WRITING	5980	12.4	17.1
ACCORDING TO MEDIUM			
PRINTED	33199	69.0	95.3
RECORDED	14758	30.6	42.3
ACCORDING TO VARIETY			
ROTE	13690	28.4	39.3
INCREMENTAL	0	.0	.0
VARIATIONAL	125	.2	.3
OPERATIONAL	33704	70.0	96.7

TABLE 11.- (b)

FRENCH THROUGH PICTURES — REPETITION, DISTRIBUTION

ACCORDING TO SKILL	NUMBER OF TOKENS	PERCENT OF TOTAL	PER HUNDRED TEXT WORDS
TOTAL WORDS OF METHOD	48144	99.9	176.4
LISTENING	20314	42.1	74.4
READING	26951	55.9	98.7
SPEAKING	28296	58.7	103.6
WRITING	14432	29.9	52.8
ACCORDING TO MEDIUM			
PRINTED	26951	55.9	98.7
RECORDED	20314	42.1	74.4
ACCORDING TO VARIETY			
ROTE	18651	38.7	68.3
INCREMENTAL	7753	16.1	28.4
VARIATIONAL	1545	3.2	5.6
OPERATIONAL	13332	27.6	48.8

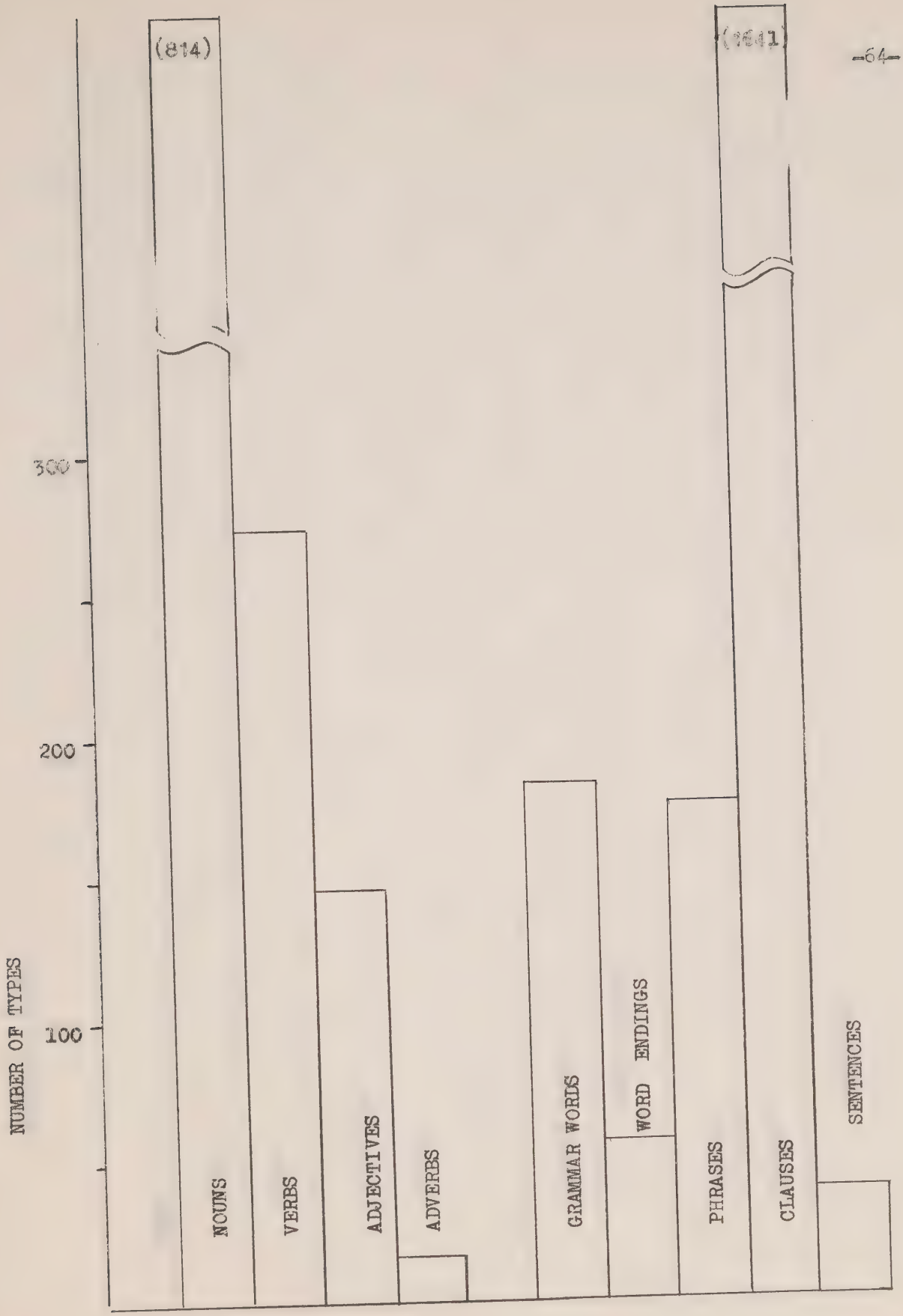


FIGURE 2.- (a) QUANTITY OF SELECTION

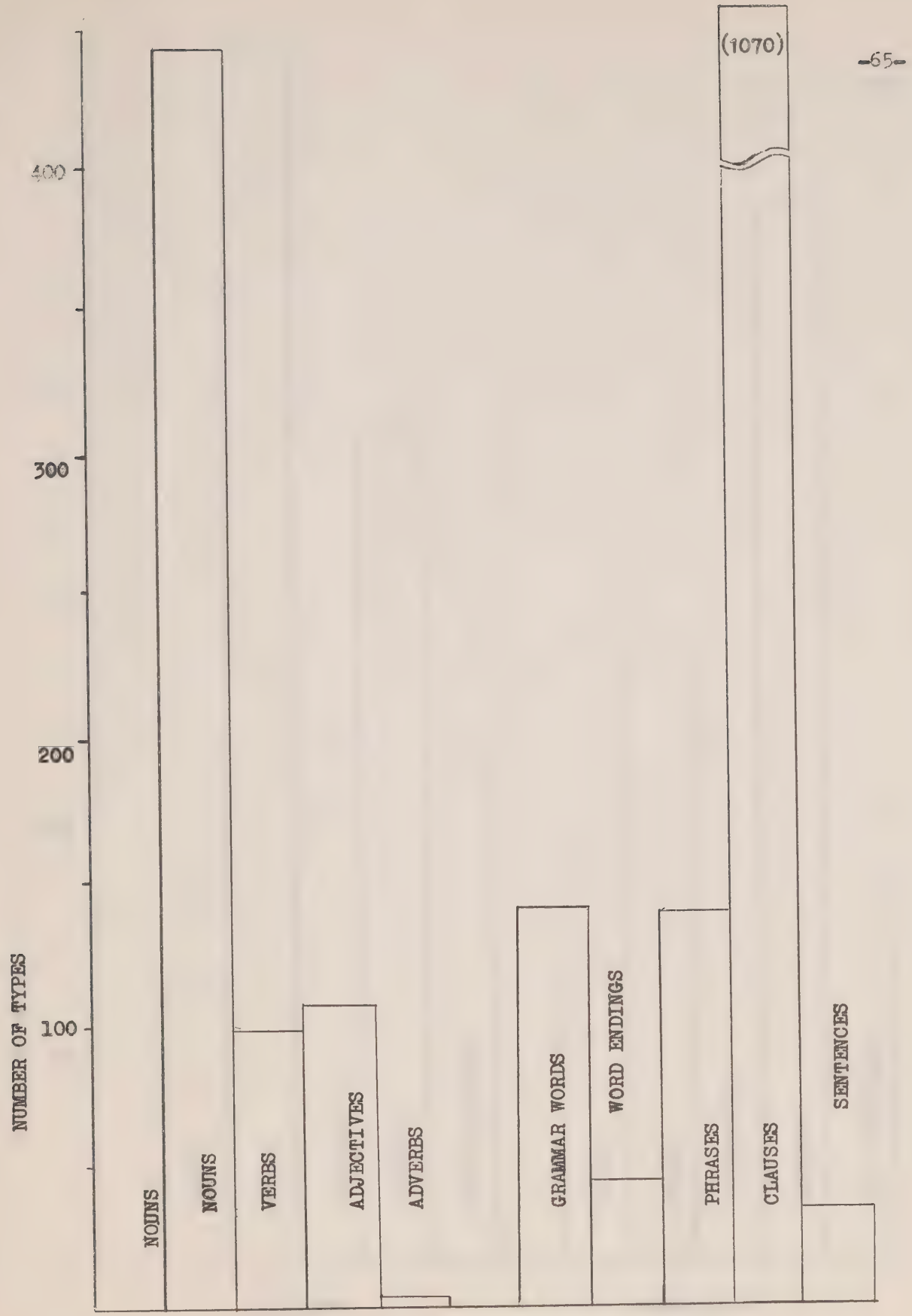


FIGURE 2.- (b) QUANTITY OF SELECTION

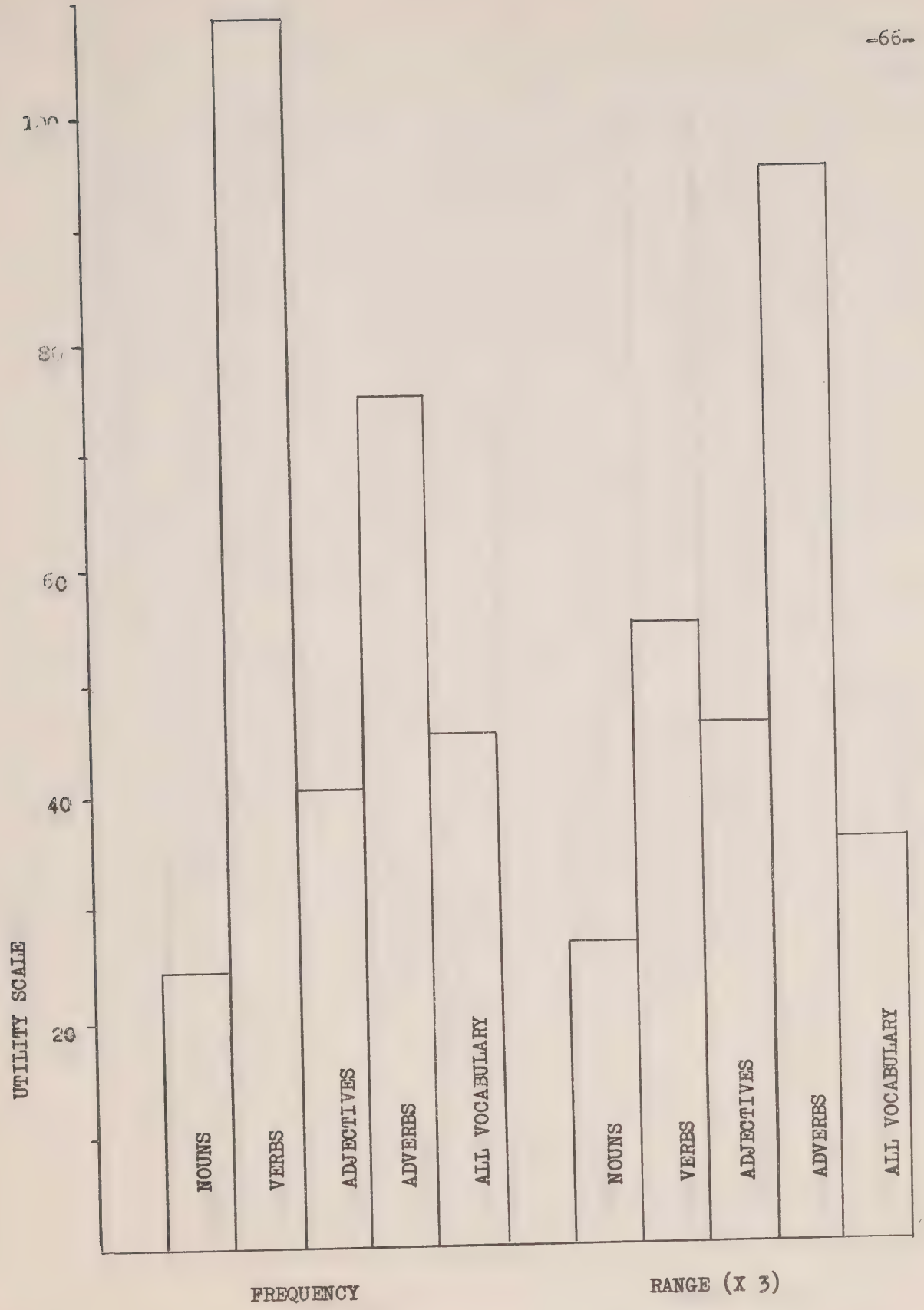


FIGURE 3.- (a) AVERAGE UTILITY

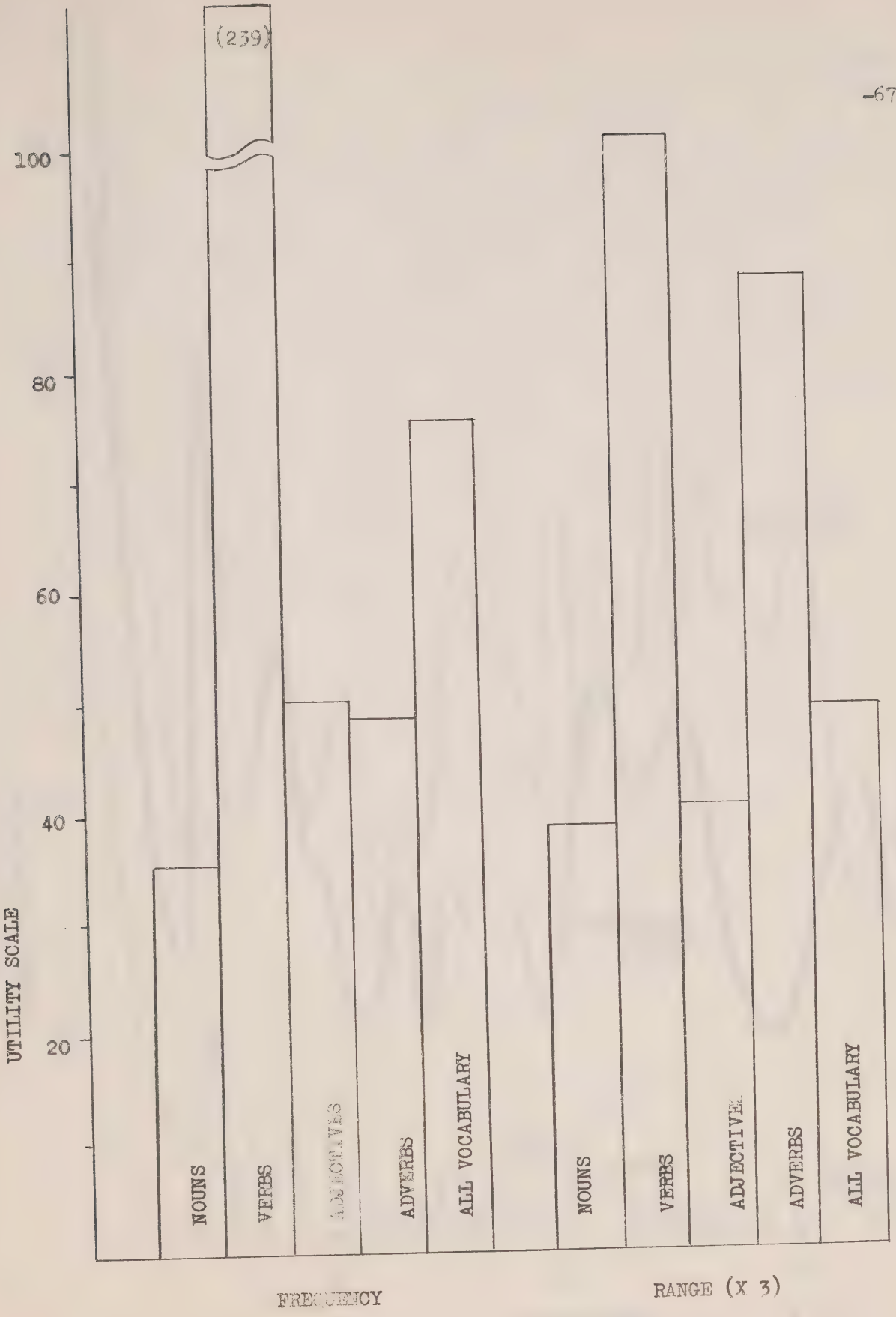


FIGURE 3.- (b) AVERAGE UTILITY

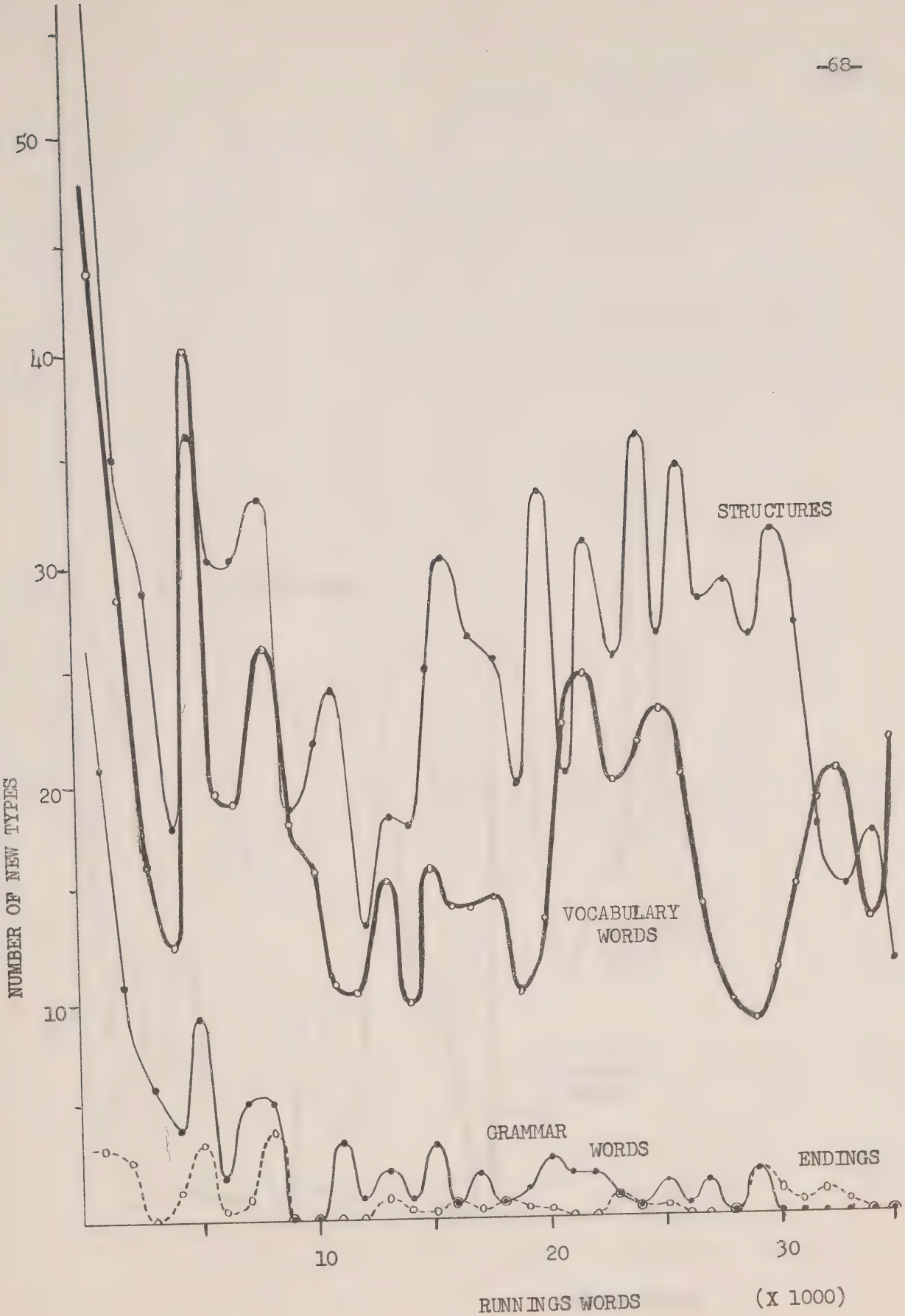


FIGURE 4.- (a) INTAKE BY 1000-WORD BLOCK

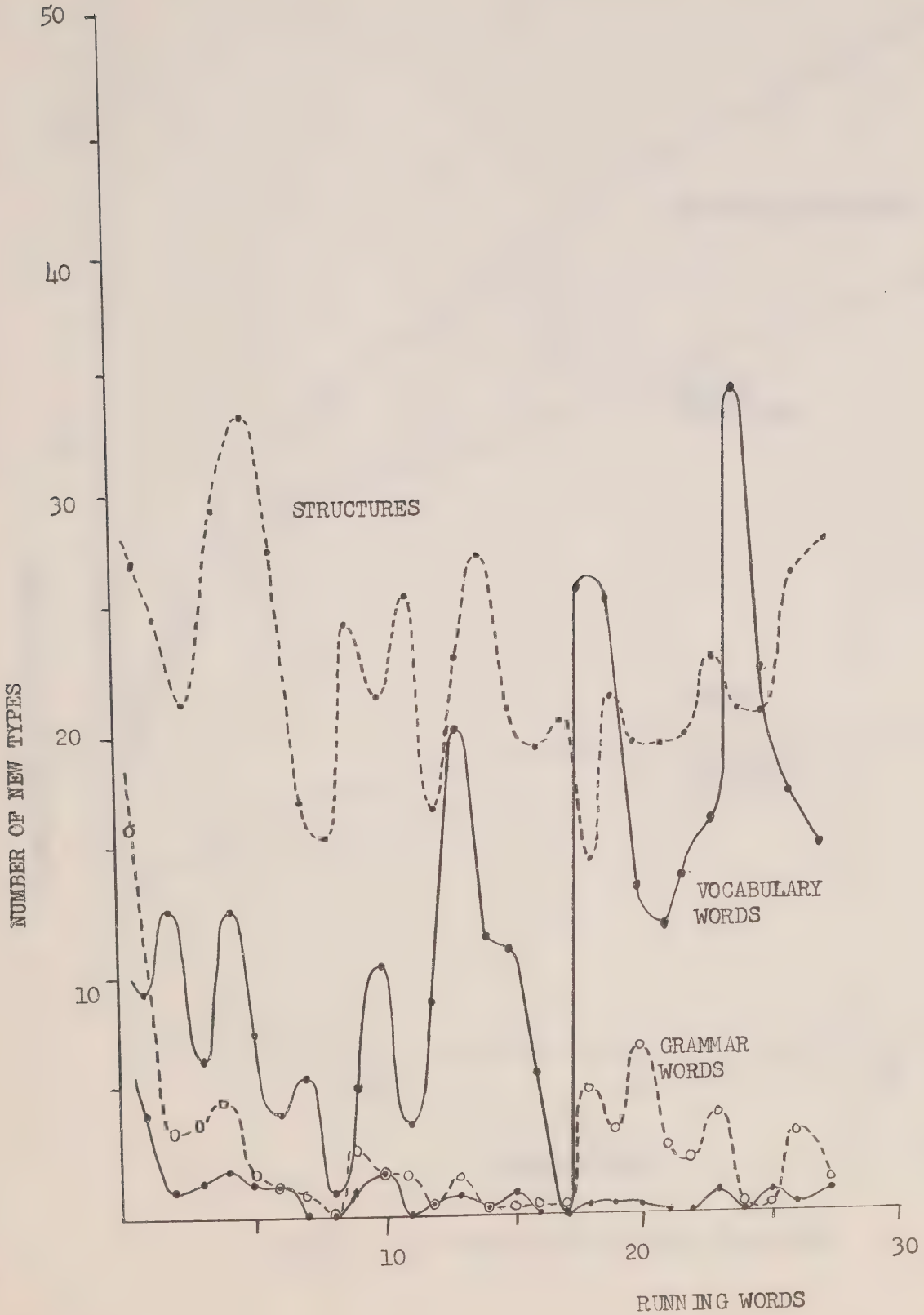


FIGURE 4.- (b) INTAKE BY 1000 - WORD BLOCK

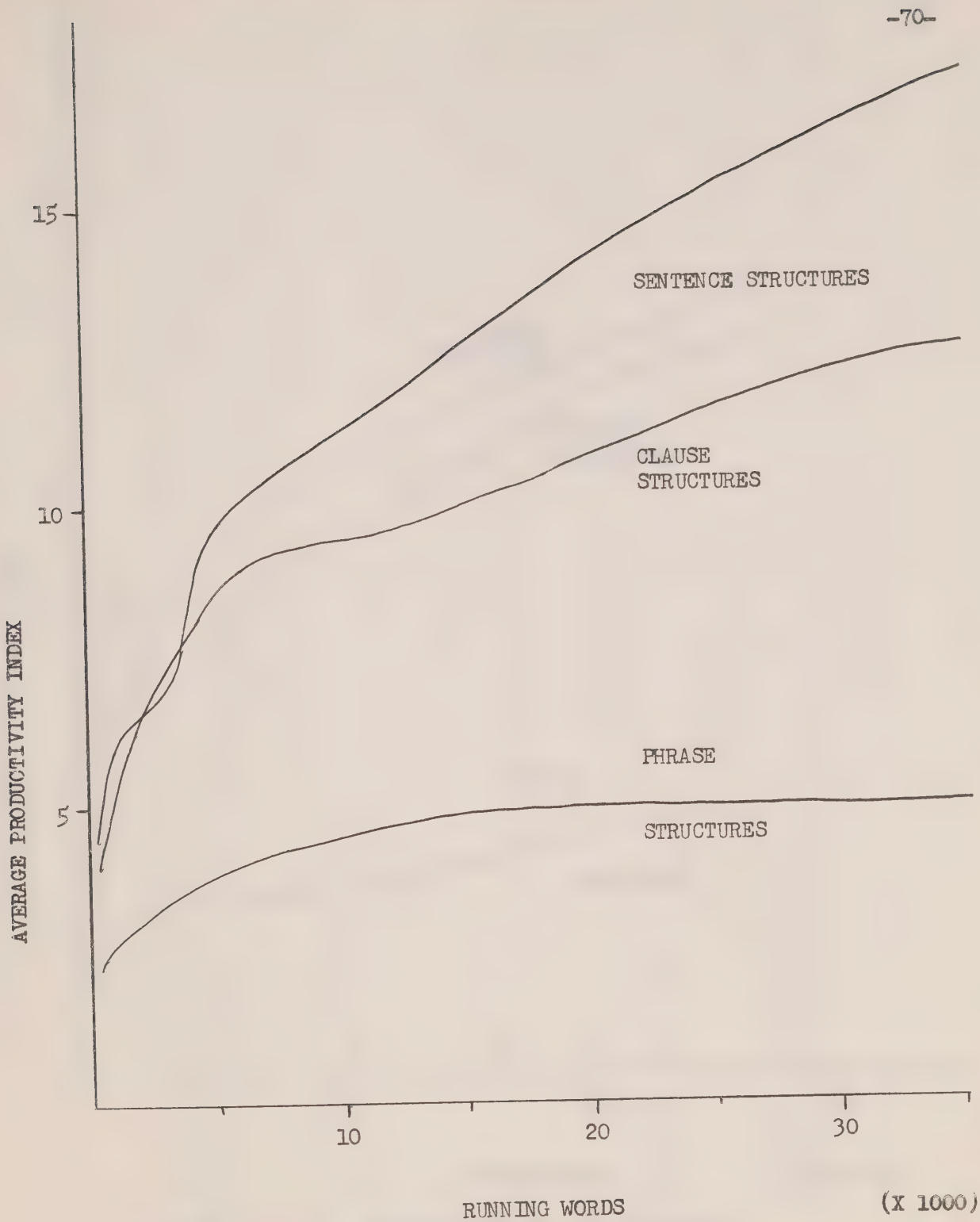


FIGURE 5.- (a) PRODUCTIVITY OF THE STRUCTURES

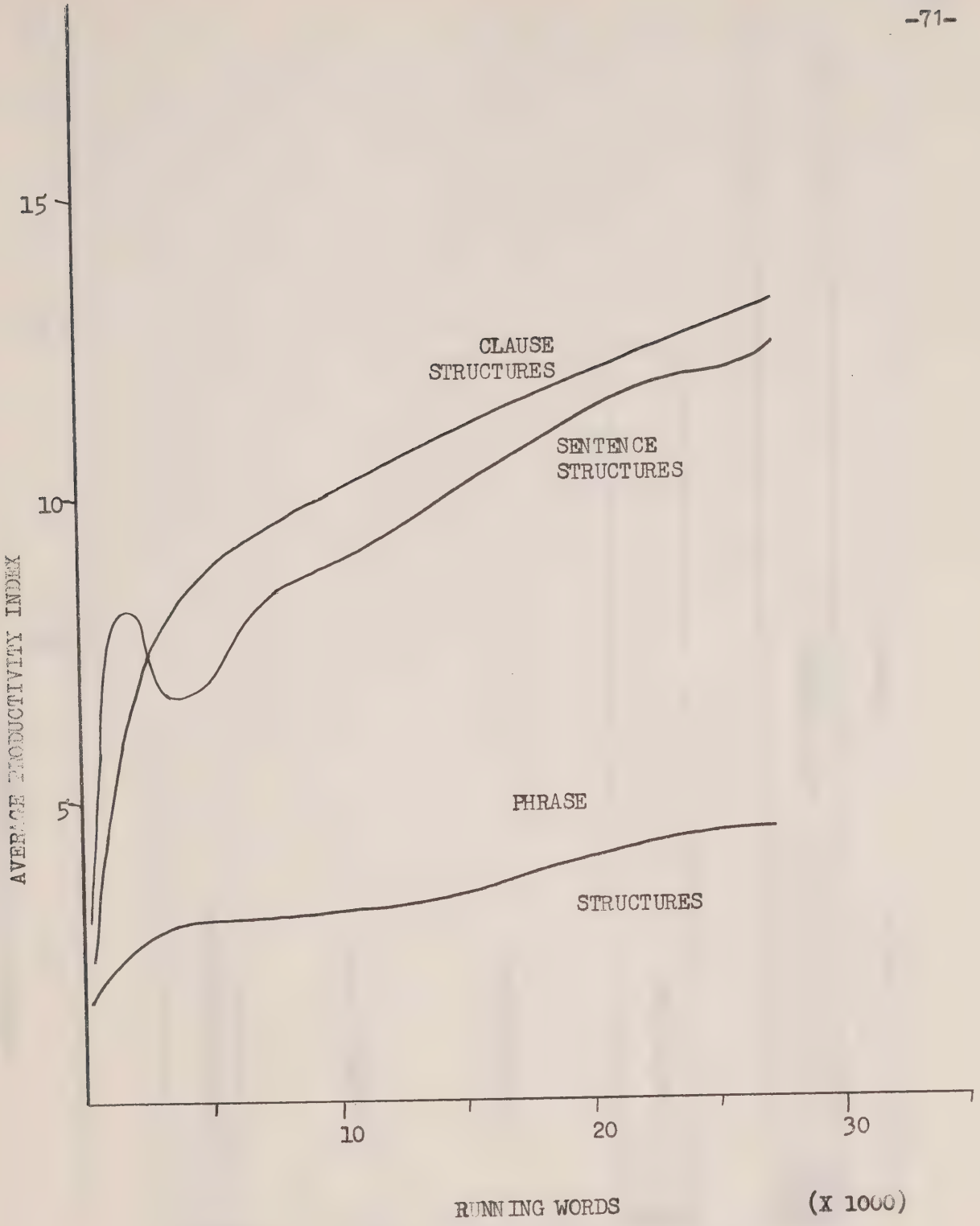


FIGURE 5.- (b) PRODUCTIVITY OF THE STRUCTURES

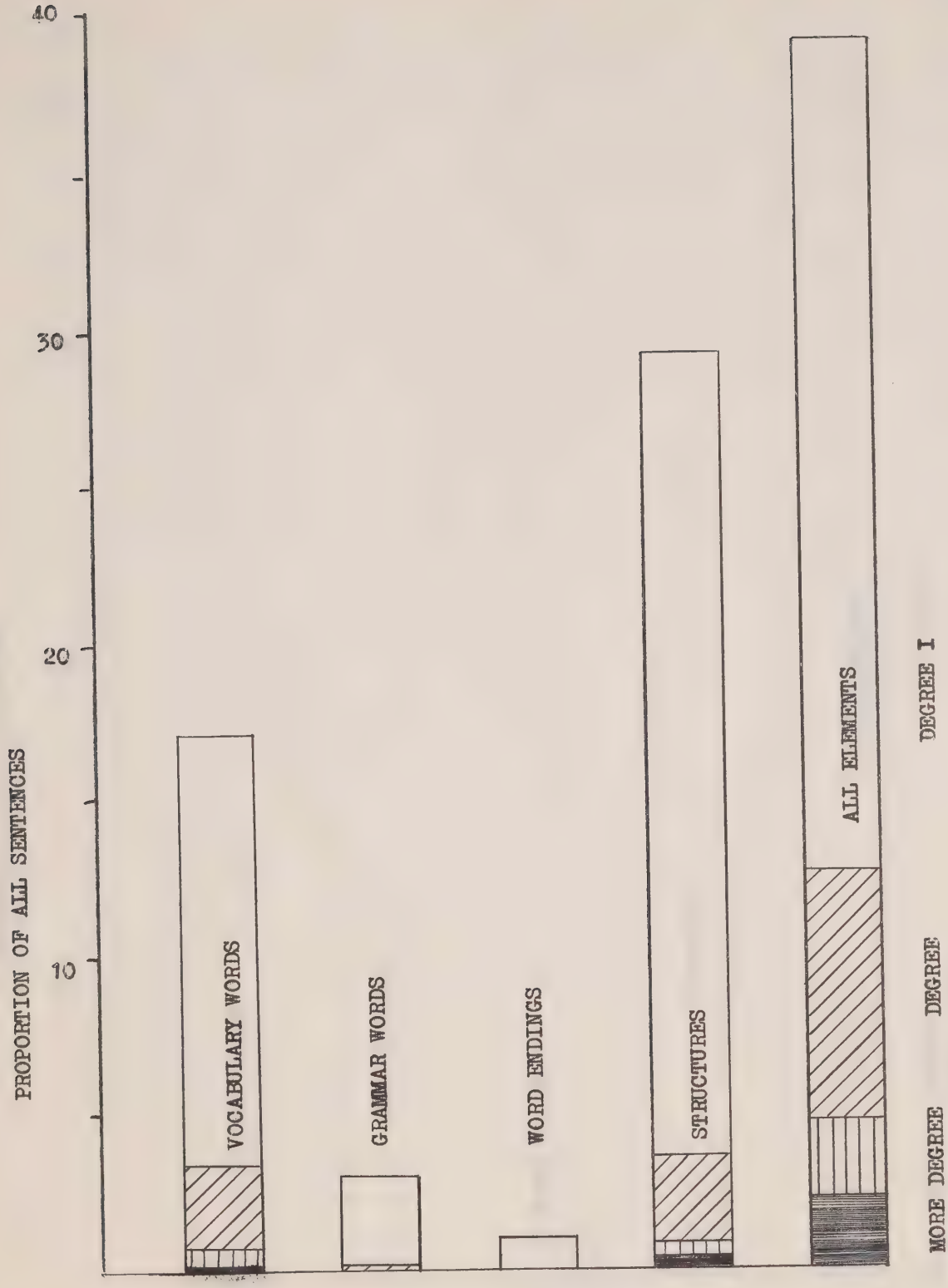


FIGURE 6.- (a) DENSITY BY SENTENCE

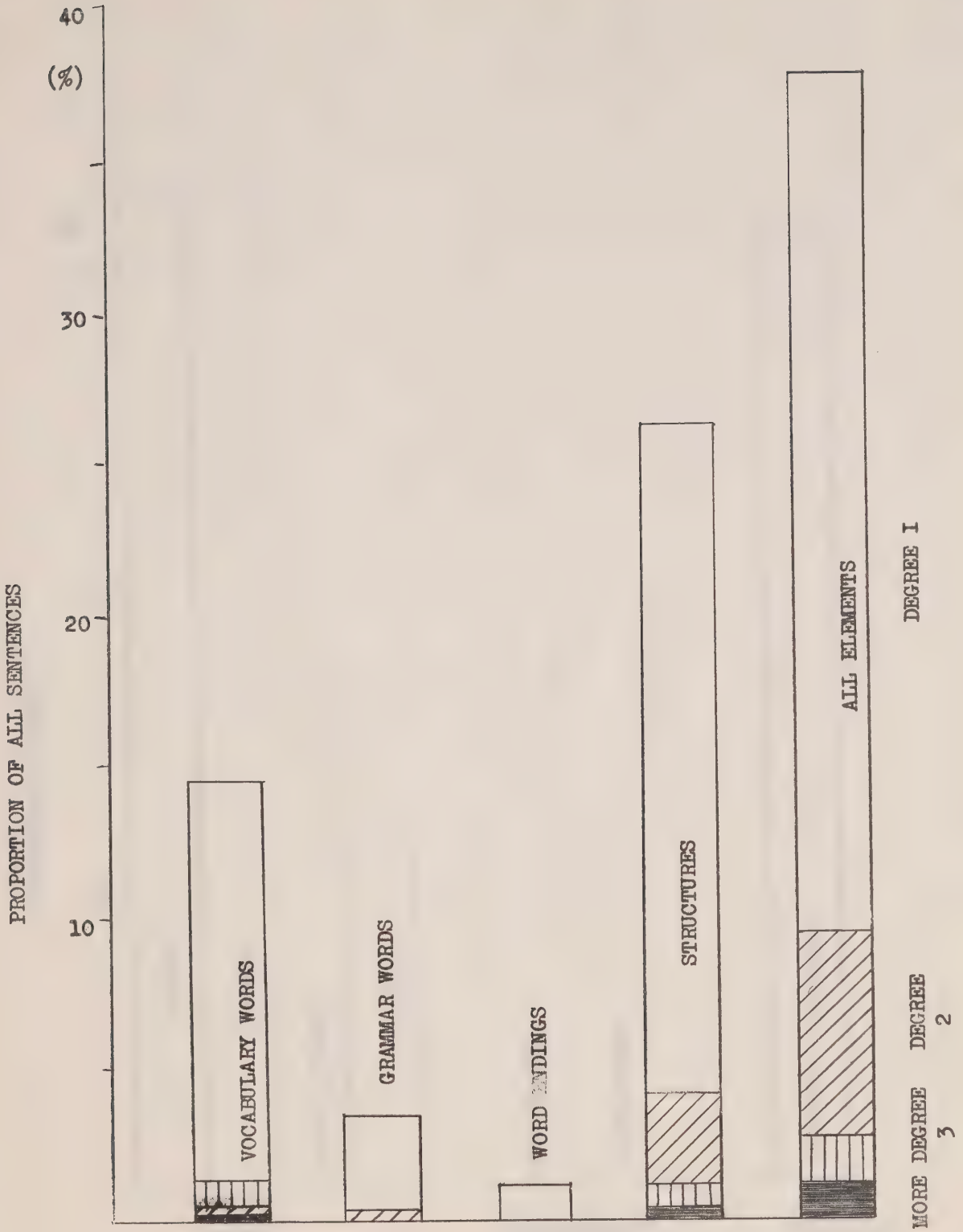


FIGURE 6.- (b) DENSITY BY SENTENCE

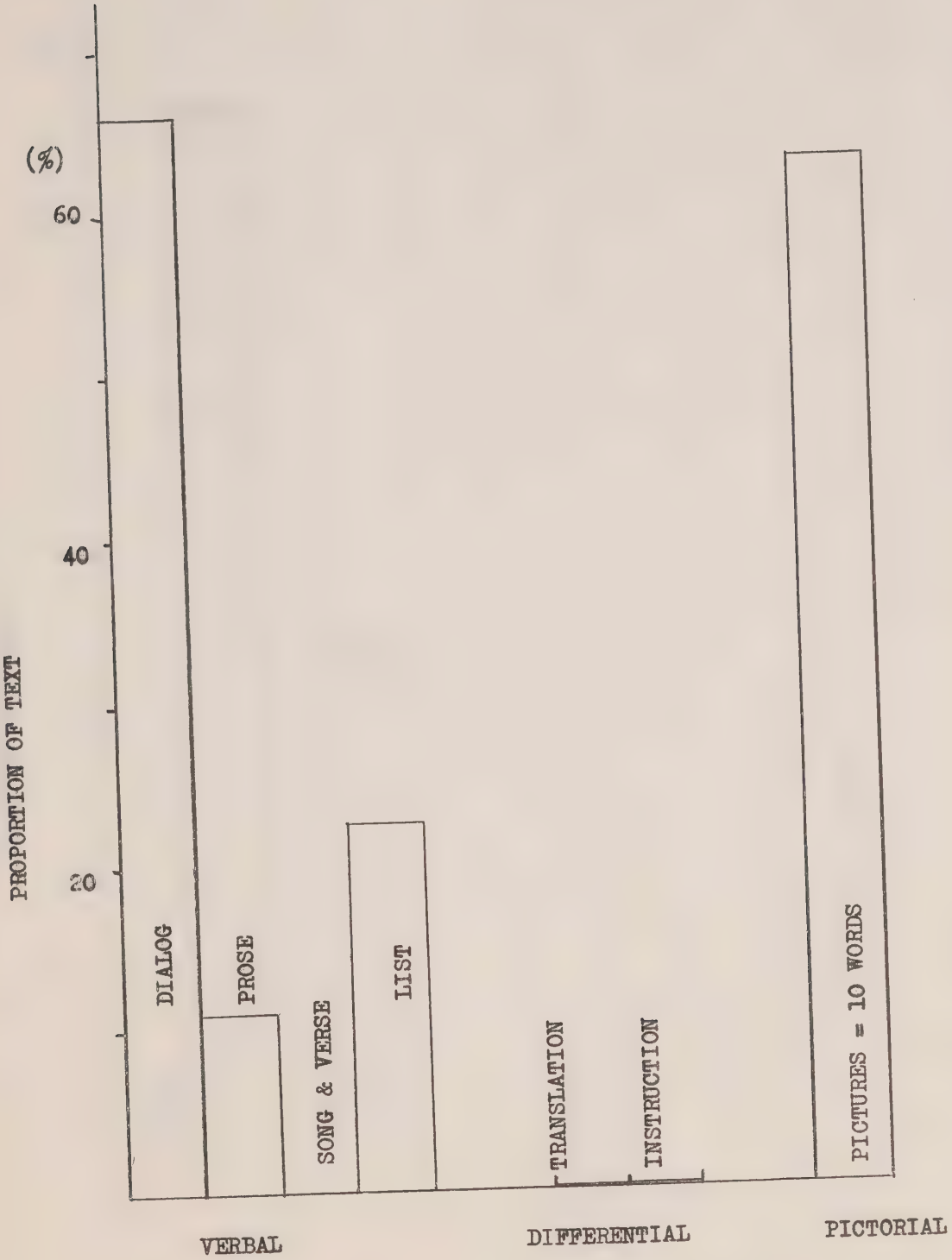


FIGURE 7.- (a) PRESENTATION BY CONTEXT

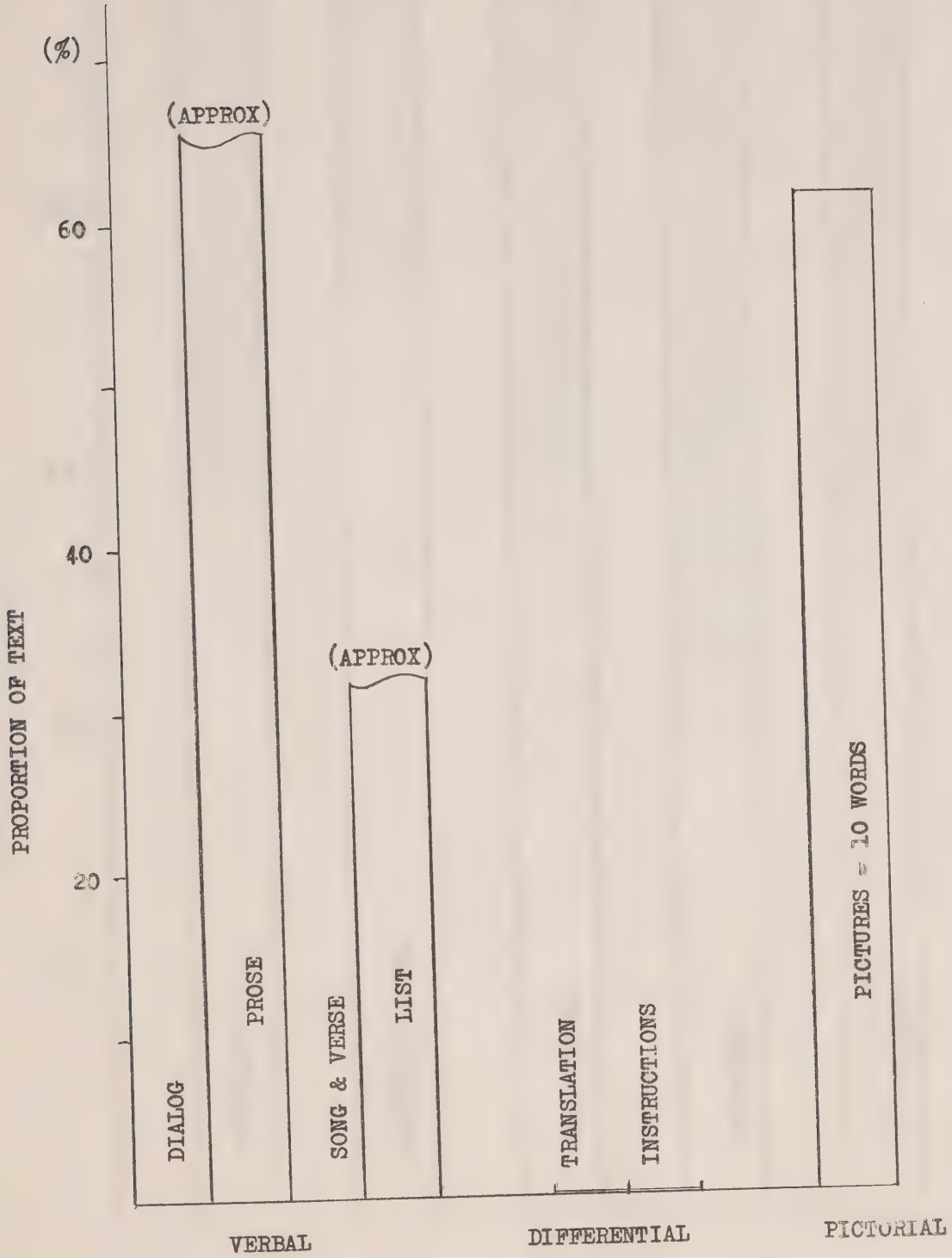


FIGURE 7.- (b) PRESENTATION BY CONTEXT

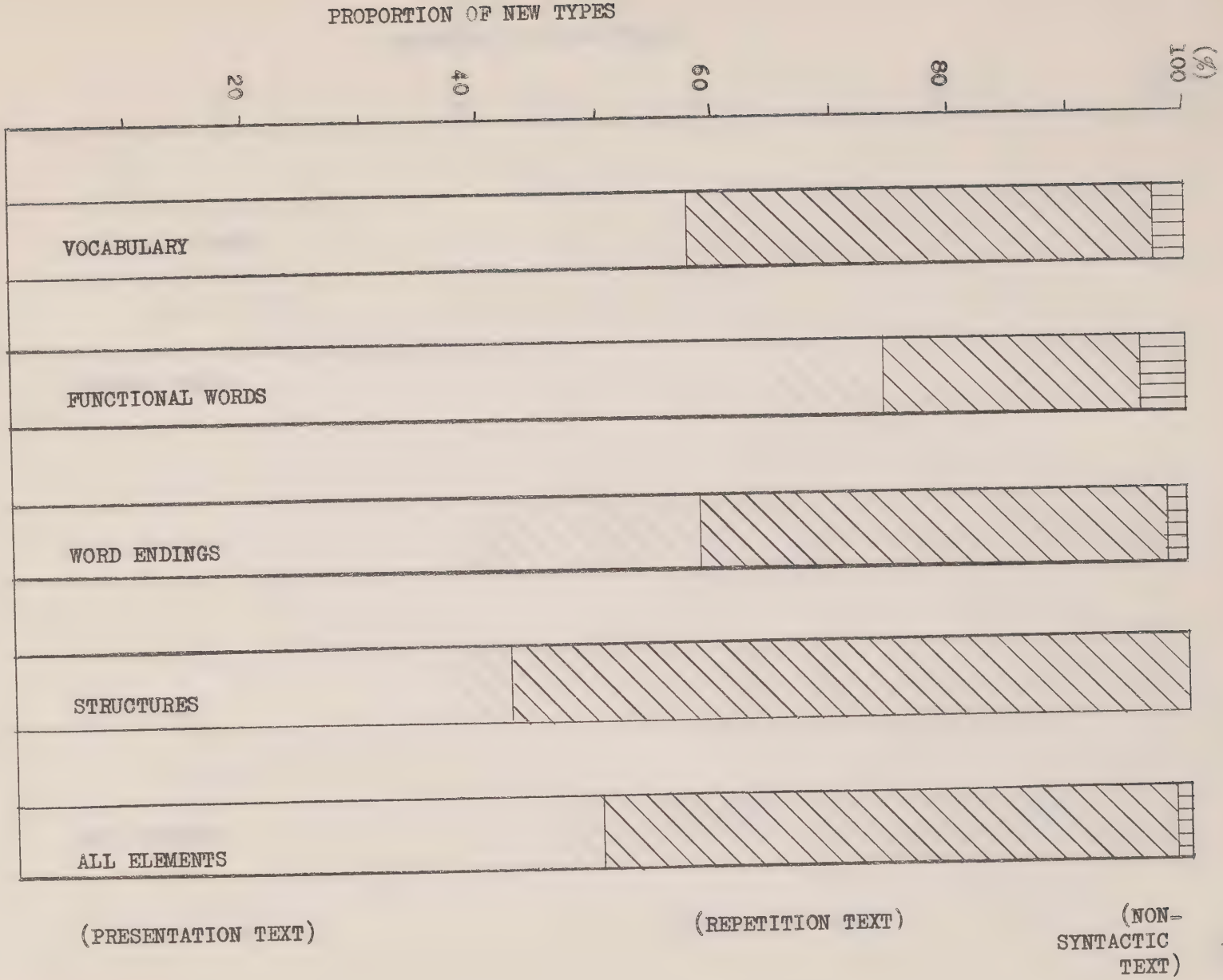


FIGURE 8. (a) PRESENTATION: INTRODUCTION

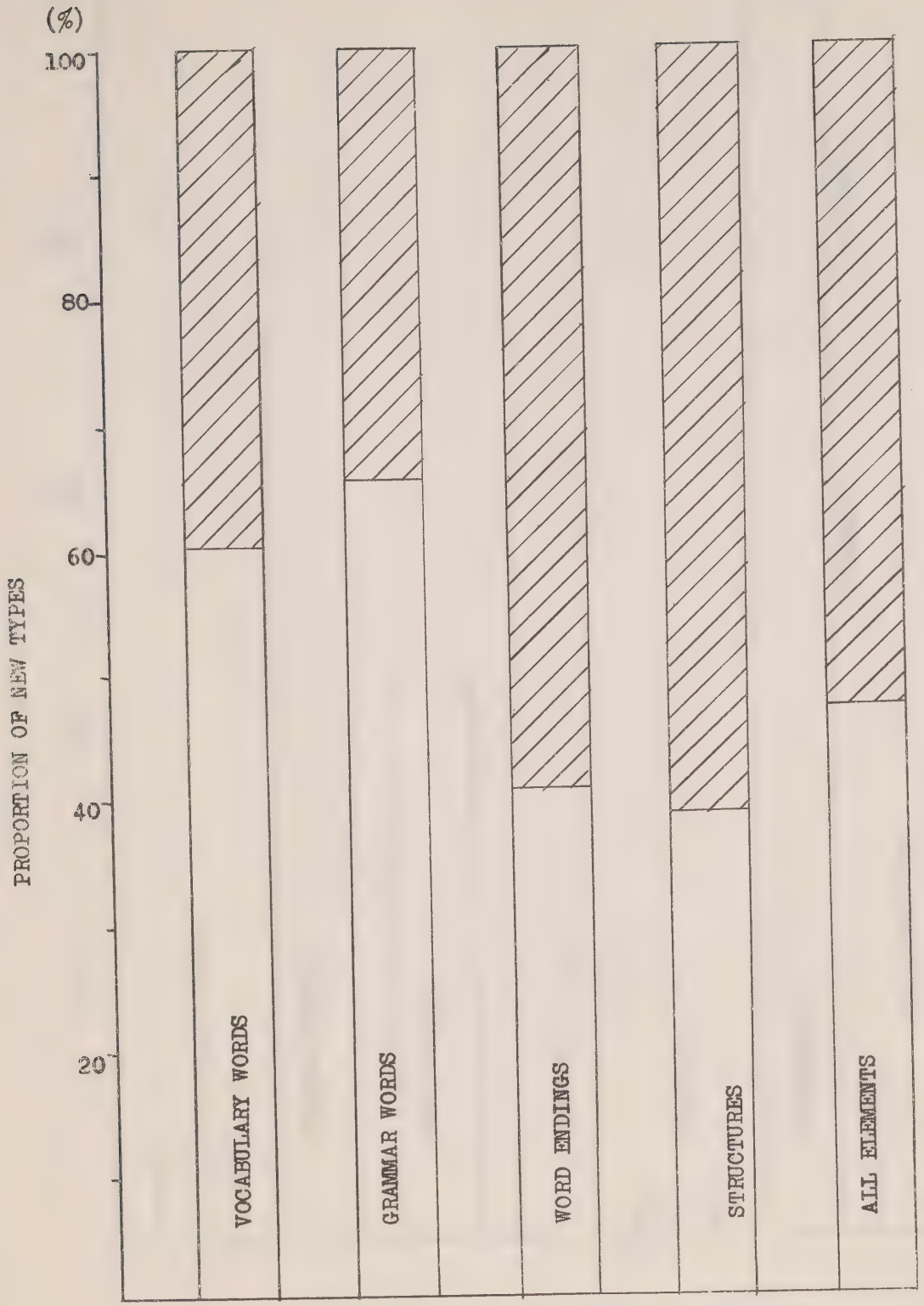


FIGURE 8.- (b) PRESENTATION: INTRODUCTION

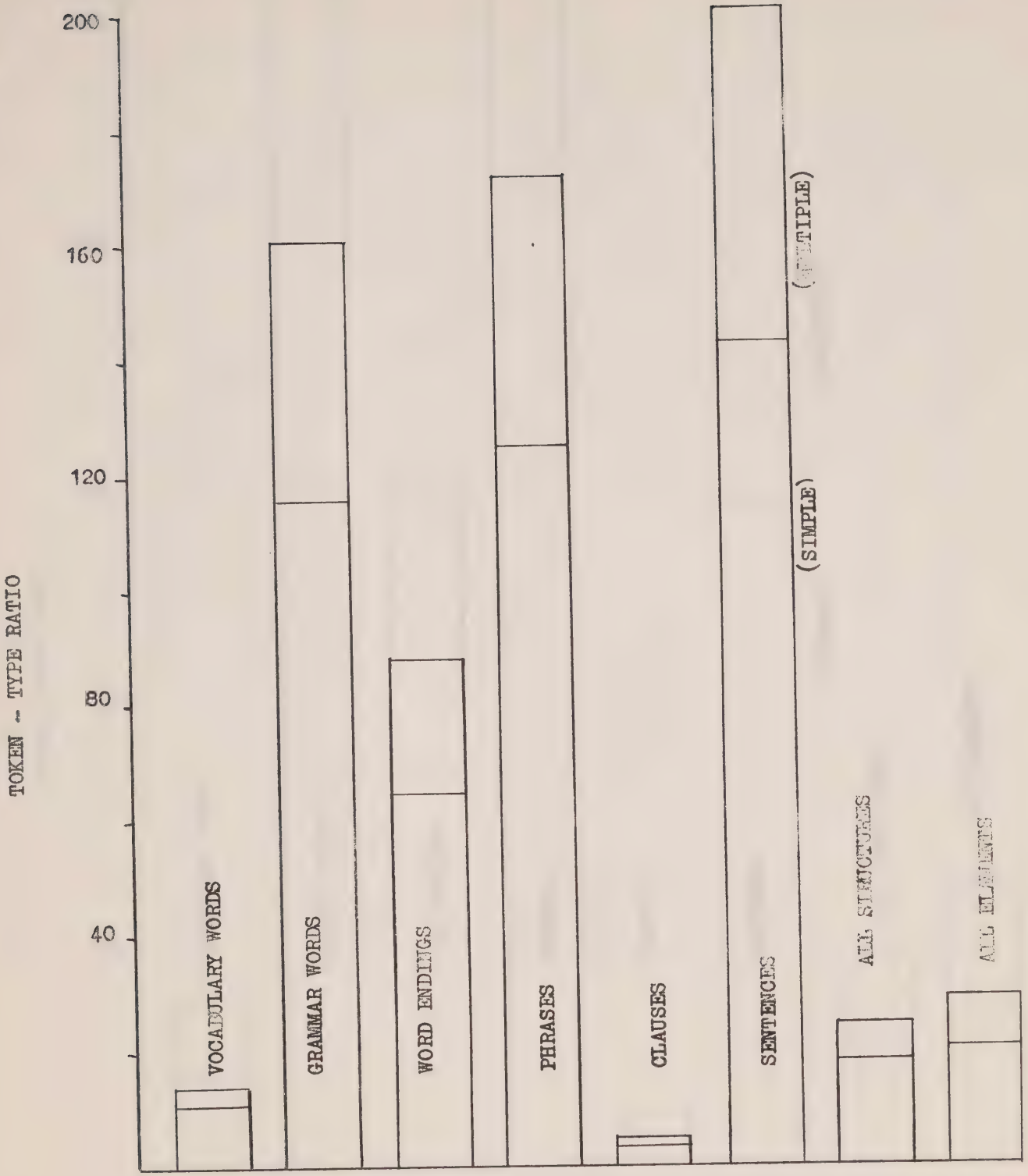


FIGURE 9.- (a) REPETITION BY CATEGORY

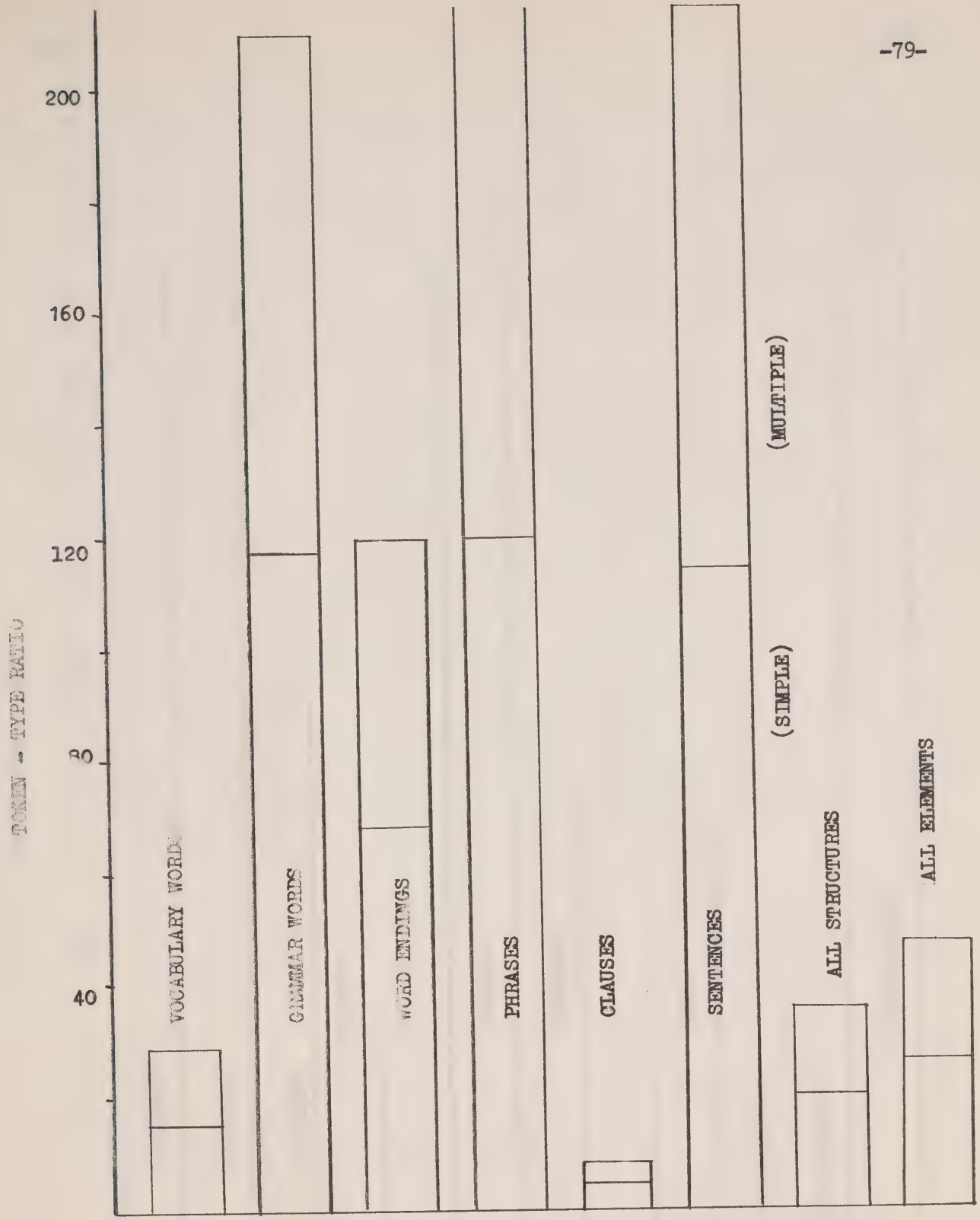


FIGURE 9.- (b) REPETITION BY CATEGORY

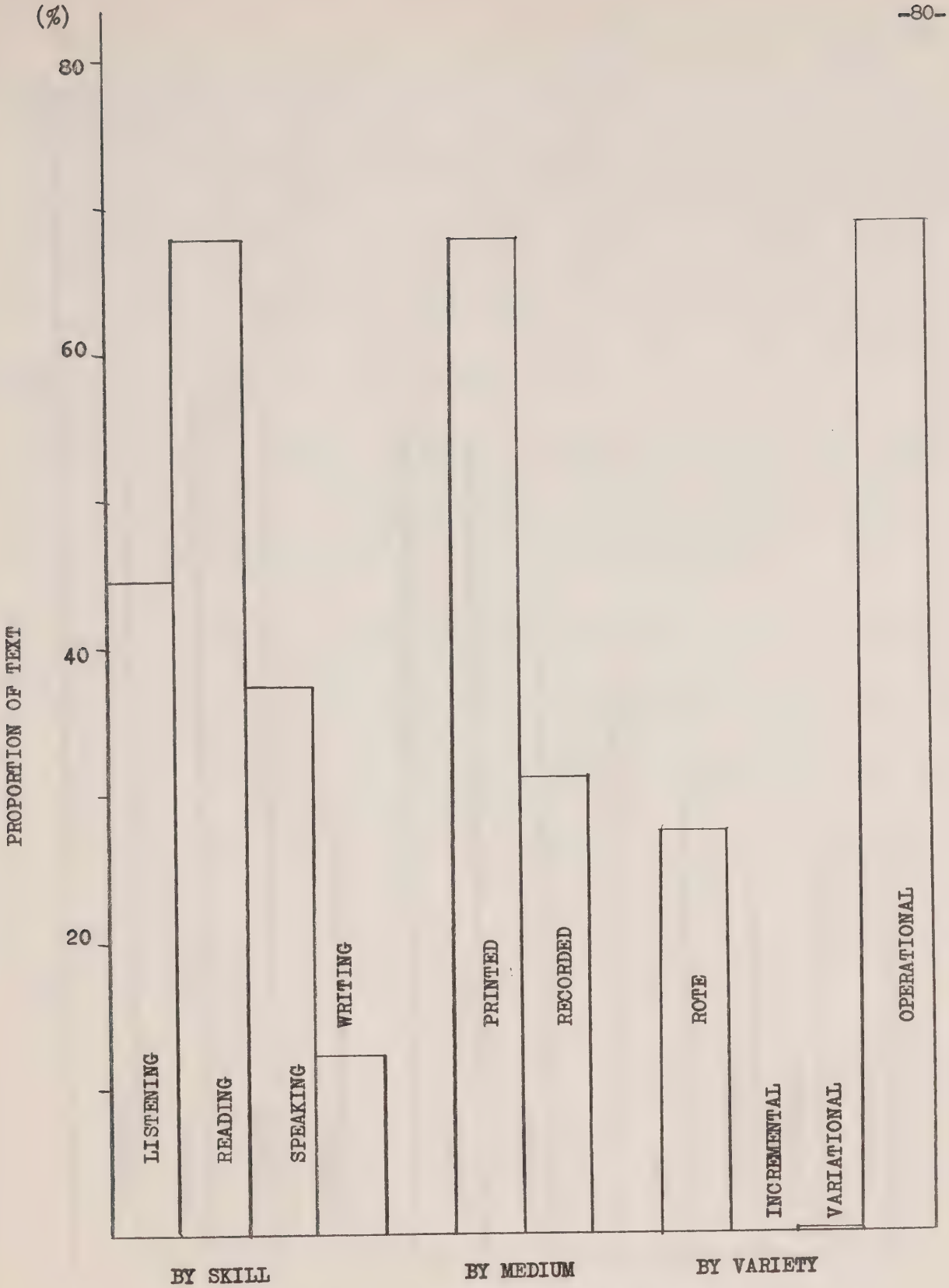


FIGURE 10.- (a) DISTRIBUTION OF REPETITION

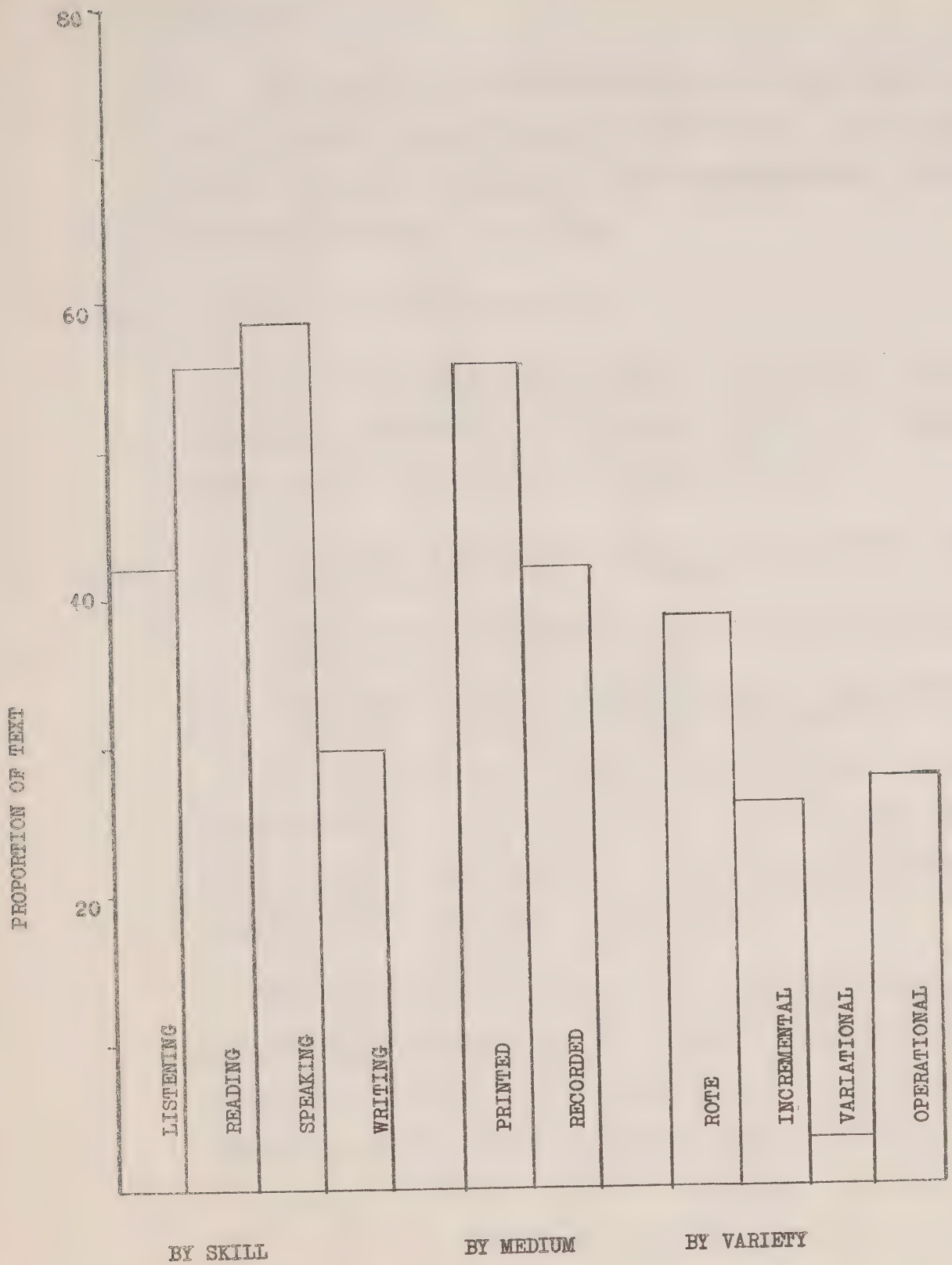


FIGURE 10.- (b) DISTRIBUTION OF REPETITION

9.- Conclusions

The evaluation of the methods in terms of their pedagogical factors is beyond the immediate scope of this project. We can, however, assess the numerical results as to their appropriateness, and the system of analysis as to its value.

9.1.- Definition of the Measurements

The intake as defined is too complex a variable to be readily meaningful. Redefining it as the ratio of new types to running words within a segment of text has the following advantages:

1. the order of magnitude of the value is independent of the arbitrarily chosen length of segment,
2. the values are independent of the repetition, or number of tokens corresponding to the types,
3. the results are more easily interpreted in terms of gradation: the higher the value, the steeper the gradation.

The total productivity is another complex variable. It depends on the number of structures, the lengths of the structures, and the number of types of each element of the structures. Averaging eliminates the effect of the number of structures.

The non-logarithmic productivity is so heavily weighted by the most productive structures that the results lose significance for the structures as a whole. A logarithmic function reduces the relative importance of larger productivity values, making the average productivity more representative of all the structures. By choosing a logarithmic base other than ten, discrimination between high and low productivities may be varied.

9.2.- Proposed Measurements

The system of analysis could be adapted to produce additional results with only minimal modification.

9.2.1.- Phonetic Description

The selection, gradation and repetition of the phonetic elements of the words is feasible. To the data files "Vocabulary Words" and "Functional Words" would have to be added the coded pronunciation of each word. With a few additional instructions the existing programs would suffice.

9.2.2.- Utility Parameters

Other criteria such as availability and semantic coverage could be used to assess the vocabulary. The appropriate values would be assigned to the entries of "Vocabulary List".

9.2.3.- Repetition Profile

In addition to the amount of repetition of an element it would be possible to assess the distribution of its repetition throughout the text. The profile could be expressed by a vector of several values, each representing the repetition within a certain proximity of the first occurrence. The profiles could be averaged for the whole text.

9.2.4.- Detailed Results

The results presented by grammatical category could be augmented to include further subdivisions of the grammar words, word endings and

structures.

9.3.- The Presentation of the Results

A number of modifications in the format and grouping of the results would add to their usefulness.

9.3.1.- Word Type List

The list "Word Types" could be modified to include only one inflectional form for each word. The word's repetition count would then be more significant. The list would be shorter, hence easier to use.

9.3.2.- Type Evaluation

The list "Word Types" could easily include the values of utility for each vocabulary word. Similarly, the list "Structure Types" could include the productivity. This would enhance their usefulness for detailed examination of the selection.

9.3.3.- Averages and Dispersion

We made frequent use of arithmetic means to summarize long lists of figures. These averages would be more useful if complemented by some measure of the dispersion of the values about the mean.

9.4.- The System of Analysis

Any modification of the system of analysis that minimizes manual intervention, saves production time or improves the output results is an improvement. A certain number are envisaged.

9.4.1.- Pre-Editing

The identification of the pedagogical units of text (ref. 4.1.) with coded sigla should not be included in the pre-editing stage. The sigla should not be included in the text, but in an independent card deck. The analyst could select prepared sigla cards, indicating on each the order number of its position of the text. The results of the sigla analysis would then be collated with the text at a late stage in the analysis. This procedure offers several advantages.

1. It eliminates handwritten entries in the original text. The key punch operators work with a simpler, cleaner text, hence, make fewer errors.

2. The proposed procedure is flexible. The sigla need not be coded before the punching up. Sigla errors are not irretrievably included in the textual data at "P" as is presently the case.

3. It is more economical. Clerical help can perform the simplified pre-editing. Coding and punching are speeded up. The text need not be scanned during the first machine pass to extract the sigla.

9.4.2.- The Word Lists

An extension of the list "Vocabulary Words" would reduce the manual intervention and improve the measurement of the utility (ref. 2.1.3.) The number of words to be identified during the manual correction (Fig. 1.- (d)) would be reduced. The correction time could easily be cut in half by doubling the length of the data list. With the longer list, the utility parameters would be applied to a larger proportion of the words.

9.4.3.- The Elimination of Functional Ambiguities

The use of a special grammar to complete the identification of homographic elements is awkward and time consuming. Each word must be tested for an ambiguity indicator, and, if positive, for a series of possible context conditions. To be completely effective, the list "Ambiguity Rules" would have to be much longer than at present.

A possible solution would be to include the neighbouring elements with each word listed as being ambiguous. During word correction, the assigned identification can be verified and adjusted in terms of each context encountered. This procedure could easily include the reduction of idioms to the word level.

9.4.4.- The Grammatical Analysis

The present grammatical analysis is sufficient, but not ideal. As employed, the three levels of structuration, phrase, clause and sentence, do not give a compact description of the grammar. That is, there are too many different structural types at the clause level. For instance, of 1855 structure types in the example of Table 3.- (a), 1641 are clause structures.

One solution is to redefine the phrase structures to include sequences of two or more dependent phrases. A series of prepositional phrases, for instance, would constitute one phrase. This would increase the number of phrase types and decrease the number of clause types. For the chosen example, a total of 950 structures, with

about 600 phrase structures and 300 clause structures would render the results more tractable.

A second solution is the inclusion of a phrase-clause interlevel. This fourth level would include the grouped phrases relegated to the phrase level in the above solution. The total for the example might be about 500 structures: 40 sentence structures, 250 clause structures, 60 groups of phrases, and 150 phrase structures. It should be possible to condense the description even further, increasing the number of phrase groups at the expense of the number of clauses.

9.5.- Economic Considerations

The practical feasibility of analysing methods depends on development of the system can be broken down into the analysis of the problem, the preparation of the data, and the programming of the machine. The cost in human terms cannot be broken down, as these three aspects were intermingled. Globally, two person-years were applied to the development of the system. More time is needed to perfect it.

Production costs vary from one method to another, depending on length and complexity. A rough break-down is given below in terms of the minimal cost and a cost-per-word supplement.

1. pre-editing: .07 cents/word
2. card punching: 1.25 cents/word
3. machine time:$75 plus 1.00 cents/word
4. coordination and correction: $20.00

5. preparation of graphs: $20.00

The total amounts to $115 basic cost, plus 2.32 cents per word. For example, a method with 30,000 words of text would cost about $800.

The cost of machine time depends on the installation. It is hoped that machine cost will be reduced at least by half with the impending installation of a more advanced machine.